AD NUMBER

AD354604

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to DoD only; Administrative/Operational Use; DEC 1962. Other requests shall be referred to Rome Air Development Center, Attn: EMIIT, Griffiss AFB, NY 13441-5700.

AUTHORITY

AFRL/IFOIP ltr, 4 Jan 2007

AD NUMBER
AD354604
CLASSIFICATION CHANGES
TO unclassified
FROM confidential
AUTHORITY
AFRL/IF Rome, NY ltr, 28 Aug 2006

AD NUMBER
AD354604
CLASSIFICATION CHANGES
TO confidential
FROM secret
AUTHORITY
31 Dec 1974 per DoDD 5200.10 document marking

CONFIDENTIAL

[redacted]

AD 354604L

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA

[redacted]

~~CONFIDENTIAL~~

AD354604

DECLASSIFIED

UNCLASSIFIED

FACT CORRELATION FOR INTELLIGENCE ANALYSIS. VOLUME 1. APPLIED RESEARCH PLAN (U)

FEDERAL ELECTRIC CORP PARAMUS NJ

15 DEC 1962

Distribution authorized to DoD only; Administrative/Operational Use; DEC 1962. Other requests shall be referred to Rome Air Development Center, Attn: EMIIT, Griffiss AFB, NY 13441-5700.
NOFORN

Declassified by:
AFRL\IF
Donald W. Hanson
Director, Information Directorate,
Declassified on: August 25, 2006

Notice
There has been a classification change to this document. It is the responsibility of the recipient to promptly remark it to indicate change.

DECLASSIFIED

UNCLASSIFIED

CONFIDENTIAL

SECURITY NOTICE

Special Handling Required. Not Releasable to Foreign Nationals.

CLASSIFICATION NOTICE

This document is classified SECRET because it discusses in detail the limitations inherent in techniques for manipulating data available to present intelligence system environments and implies a course of technique development for future application to intelligence data handling environments.

AVAILABILITY NOTICE

Military agencies may request copies from DDC, all other qualified requests must be submitted through RADC (EMIIT), Griffiss AFB, NY.

PATENT NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

SECRET

RADC-TDR-62-461, Vol. I

15 December 1962

Final Report

(U) FACT CORRELATION FOR INTELLIGENCE ANALYSIS

Volume 1: Applied Research Plan

ITT International Electric Corporation

Route 17 and Garden State Parkway, Paramus, New Jersey

A SUBSIDIARY OF INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

Technical Report P-AA-TR-(0007)

Security No: IEO 62-P-68

Copy No. 36 of 60 Copies

Contract No. AF 30(602)-2739

Project No. 4594, Task No. 459403

Prepared for

ROME AIR DEVELOPMENT CENTER

Air Force Systems Command

United States Air Force

Griffiss Air Force Base New York

FOREWORD

This volume is the first of two volumes in a report analyzing the requirements for an applied research program for an automated system capable of correlating facts for intelligence analysis. This volume discusses the over-all research plan. (S)

This report was prepared by the International Electric Corporation (IEC)--a subsidiary of the International Telephone and Telegraph Corporation (ITT)--Paramus, New Jersey. The study leading to this report was performed under Contract No. AF 30(602)-2739 for the Intelligence and Electronic Warfare Directorate of the Rome Air Development Command, Griffiss Air Force Base, New York. (U)

This study was conducted within the Advanced Analysis Department under the direction of Jacques Harlow. The staff that performed the analytical and theoretical studies included Quentin A. Darmstadt, Dr. George Greenberg, Maralyn Lindenlaub, David M. Massie, Dr. Howard E. Smokler, Alexander Szejman, and Alfred Trachtenberg. (U)

The staff acknowledges the contribution of the authors cited as references and of the equipment manufacturers who supplied information pertaining to their present equipment design specifications and their future plans for equipment development. The report also contains original concepts developed by members of the staff while performing research activities sponsored by IEC. (U)

ABSTRACT

This report presents an analysis of the requirements for a Fact Correlation System--a specialized information storage and retrieval system oriented to the problem of correlating facts for intelligence analysis. Two aspects of the problem were analyzed: the equipment requirements necessary to implement a highly automated system, and the processing requirements necessary to transmute isolated items of information into an integral whole. The report includes a broad review and survey of existing processing techniques, primarily in the fields of linguistic analysis and adaptive learning, and a survey of both existing and experimental data processing equipment. This survey and analysis form the basis for a research plan for developing the techniques and equipment required to correlate facts automatically. (S)

This volume of the report discusses the basic concept of information retrieval; reviews the system requirements for a Fact Correlation System, including the specific functions of personnel, equipment, and programs together with their interrelationships; and develops a generalized schedule and plan for research, development, and implementation leading to the installation of the system. The conclusion derived from this study is that at least ten years will be needed to develop a fully automated system, although less sophisticated functions may be available within three years. The research plan recommends a method of approach for developing the system in a series of stages. (S)

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF ILLUSTRATIONS

LIST OF TABLES

I. INTRODUCTION

A. Purpose

The purpose of this report is to present the results of an analysis of the requirements for a Fact Correlation System--a specialized information storage and retrieval system oriented to the specific problem of correlating facts for intelligence analysis. Secondarily, the report describes a general research plan for developing the techniques and equipment for implementing such a system. (S)

Although specifically oriented to intelligence analysis, this study considered the problem of fact correlation within a generalized framework. The basic assumption underlying this study was that information is more important than documents. It follows that it is the information about an event or an item of knowledge from a set of documents that must be stored and retrieved. This concept is equally applicable to a variety of problems in information flow and decision making as well as to the specific problem of intelligence analysis. (U)

The elements of functional requirements were analyzed and are presented within the context of a system concept. This system was defined as:

> An information system is imbedded within an environment of data or intelligence information. The system consists of three operational functions--personnel, equipment, and techniques (including computer programming) together with the interactions among each function--required to interpret, rationalize, and understand communications from its environment.
>
> The purpose of an information system is to extend the performance and effectiveness of individuals, including intelligence analysts, interacting within the frame of reference of the system.

The nature of this problem is such that a definitive solution cannct be produced within a limited period of time. The primary purpose of this study, therefore, was to stipulate the research activity that would contribute to the solution of the problem. (S)

This report emphasizes two areas, techniques and equipment. The first step was to formulate a theoretical framework in order to establish a systematic procedure for reviewing available techniques. The theory pertains to the analysis of information presented in the form of language. Thus the fundamental techniques that were reviewed are linguistic transformation and adaptive learning. This theoretical framework necessarily biases and limits the review. Any critique of existing techniques or concepts should only be considered within the context of this frame of reference; this report is not intended to detract from viable research conducted for a different purpose. (U)

The analysis of techniques and equipment in their present form leads naturally to the formulation of requirements for additional research and development to attain the objectives of fact correlation. This report summarizes these requirements in a long-term applied research plan for extending the potential of information retrieval beyond the limited scope of simple document retrieval. The classical concentration upon special indexing terms, or such novel concepts as descriptors, has been dropped in favor of research into the interactions of groups of words or sentences. (U)

B. Scope

The analysis of the requirements for a Fact Correlation System are clearly limited. This study covered research that has been performed and that should be performed within the context of the conceptual framework of the system. This report does not purport to present substantive answers or a formal solution to the problem of correlating facts. (U)

The results of this study consist of two elements:

(a) A Research Program Plan - a plan for specific research ultimately leading to a Fact Correlation System. The plan includes a procedure for conducting research, a perspective of the interrelationship among tasks, an overview of program scheduling and phasing, expected results, and man-power requirements and qualifications.

(b) An Analysis of Technical Problems - the analysis of techniques, equipments, and technical problems includes a discussion of alternate approaches, the recommended approach together with a rationale for this recommendation, and a listing and discussion of basic tasks.

These elements of the study are reported in terms of what has been done, what should be done, and why it should be done. The review of past activity is critical; but the critique is bounded by the framework established for the analytic study. The concept presented for a future system is feasible; yet the conceptual system may be revised, and should be, as research differentiates between the ideal and practical limits of theory. (U)

C. Background

For nearly a decade the dual problems of linguistic analysis and information retrieval have been the subject of intensive applied research. The former was primarily oriented to the problem of machine translation; the latter, to the retrieval of documents from a filing system or library. The ultimate objective was to cope with the influx of scientific information on a timely basis. Secondarily, the introduction of automated techniques raised the hope of conserving human energy and time in performing essentially clerical functions. Finally, the possibility of machine translation promised the additional benefit of rapidly exchanging scientific information among scientists throughout the world. The military and intelligence potential of these concepts was hardly sublimated. (U)

The analysis of information retrieval problems tended to be limited to the library problem. Consequently, even the most sophisticated system is limited to the elucidation of information about a document in response to directed inquiries. Except for systems like the Minicard System, no complete documents are retrieved; nor is information, in the proper sense of the term, retrieved. The file system of index cards has been automated; Its new guize is sometimes more efficient, but not necessarily. In effect, the term information retrieval is a misnomer. (U)

This study attempts to outline the requirements for processing textual information--the contents of documents--to the maximum degree possible with automated techniques. The analysis was directed to the definition of new problem areas in advancing beyond current concepts and technology in information storage and retrieval. In effect, the preliminary aspects of the analysis were exploratory. In this respect this study is the antithesis of reports by Drs. M. Taube and Y. Bar-Hillel. Both men have contributed valuable ideas to the existing field of information retrieval; there is no need to denigrate their contribution. Yet their studies are laced with a sharp skepticism and unwarranted pessimism. There is no question that a theory remains untenable until it is formulated on a scientific basis.

But the squashing of ideas at their inception has limited and restricted the introduction of new concepts to the field of information retrieval. (U)

The principal new approach in this report is to combine the fields of linguistic analysis and adaptive learning with the field of information retrieval. The combination is not necessarily unique; the application may be, especially in conjunction with adaptive learning or a self-organizing system. The fundamental tenet of this analysis is that any question of information can only be resolved on a linguistic basis. The past distinction between research in these two problem areas has been unnatural, since they are interdependent. Nor is it sufficient to presume that linguistic problems exist only in the field of mechanical translation; only recently has it been demonstrated--by Dr. V. Yngve, for example--that a limited knowledge of linguistic phenomena presently confounds successful machine translation. (U)

Information retrieval has not been totally divorced from linguistic analysis. But the principal effort in linguistics applied to information retrieval has been concentrated upon manual or automatic forms of indexing and abstracting. This emphasis has been forced by the concept of retrieving information about a document rather than the information itself. The correlation of facts, in contrast, will attempt to <u>isolate</u> the information within a document, <u>store</u> the information, <u>combine</u> explicit and implicit relationships, <u>eliminate</u> redundancy, and, finally, to <u>retrieve</u> specific information upon demand. To perform these operations implies linguistic as well as inferential processes. Furthermore, the linguistic question implies a semantic capability. Indexing and abstracting as currently defined are no longer pertinent. (U)

In summary, this analysis of fact correlation is exploratory. The purpose of the study was to review linguistic and learning problems in the area of techniques; to review equipment problems in a tentative scheme for implementing these techniques. The functions of indexing and abstracting were sublimated (but not completely ignored). The study explored rules and principles for use in computer programs. The ultimate objective of the study was to develop a research methodology and plan to convert the concepts into an operational system for correlating facts. (U)

D. Organization of Report

This report consists of two volumes. The first contains a general analysis and description of requirements. The second describes available and exploratory techniques and equipment; it also anticipates specific problem areas. (U)

This volume of the report presents the Applied Research Plan. The volume includes a discussion of basic concepts of information retrieval; a review of system requirements for a Fact Correlation System; and a general plan and schedule for research, development, and implementation leading to the installation of an operational system for correlating facts. (U)

Volume 2 consists of a detailed analysis of techniques and equipment that are discussed generally in this volume. The functions of linguistic transformation and adaptive learning are described, and existing research is reviewed in terms of its ability to meet these functions. Programming requirements are cursorily reviewed; the major problems in this area can only be discerned after the nature of the programming tasks have been stipulated as a result of research in linguistics and learning. The review of equipment functions and capabilities evolves into a recommended system configuration. Design principles and exploratory research have also been reviewed to indicate areas of research activity that could be directed to the improvement of particular equipment functions. (U)

II. BASIC CONCEPTS

A. Foundations of Information Retrieval

The concept of information retrieval was originally limited to a single problem; specifically, the documentation problem. As the body of recorded information became more extensive and diversified, classical methods of indexing were outmoded. This situation is particularly apparent in the realm of science, where recent discoveries distort the time-honored but fixed molds for classifying subjects. Searching for information related to one of these new fields of knowledge became a burgeoning problem as the number of fields and the amount of documentation taxed the limits of conventional indexing and classification schemes. It is not accidental that the nature of this problem was first recognized by commercial research laboratories. (U)

One of the paramount problems in the literature search is to ascertain that all the information pertaining to a subject is found. The ancillary issue of gathering too much information is a selective function; as it pertains to manual systems, the problem is essentially one of discrimination. But the problem of retrieval tends to be fixed by the type of system used to classify information. The information must be within the domain of the system; a particular item of information does not exist so long as it remains outside the bounds of the system. Once obtained, however, the information must be uniquely classified as an entity and as an integral part of all similar information. If the classification of one item of information differs from the classification of a similar item, it may be impossible to relate the two. So long as this condition exists, all information pertaining to a particular subject cannot be retrieved. The diversity of information compounds this problem. (U)

The original research in information retrieval was thus focused upon the question of searching for the source of information. The crux of the problem appeared to be the indexing or classification systems used to indentify all pertinent documents. The salient problem was resolved. The outstanding contribution of this research was the creation of open systems of classification or indexing; a set of boundaries established *a priori* could no longer restrict the range of references to pertinent information. Yet a major difficulty persisted: an open classification system implies that all terms are independent, even for other citations of the same term, except for the exclusive set of terms for a single document. (U)

The value of this recent research is often unnecessarily castigated. It is easy to denigrate these new techniques, but usually for the wrong reason. Most of the concepts have been formulated since World War II. At first their application was oriented to manual systems such as card indexes or sophisticated systems such as Termatrex. But almost simultaneously there was a concerted effort to automate the process of literature searching. Since conventional systems were too inflexible or obviously unworkable, the new manual concepts were transmuted into either mechanical or electronic schemes of automation. Critics immediately recognized deficiencies. But these deficiencies arise from organizational errors, not conceptual errors. The Uniterm concept is still a pragmatically useful idea for a manual system; in such a system discrimination is a function of a human being. If the concept falls short of expectations in an automated system, the fault lies in the application of automation. In reality, a new concept was necessary when storage and retrieval was assigned to machines, since discrimination has become a function of the automatic processes. (U)

The influx of automated processes introduced a secondary dimension to information retrieval. The conventional concept was satisfied with an adequate description of information to indicate its location. No information per se was retrieved--the nature of the function is evident in the term "literature searching." It was sufficient for an indexing system to identify the storage place of a document. (The cogent question of information was irrelevant!) In the design of an automated system, however, an immediate issue became the storage of indexing data, the essential function of file cards. Mechanical systems based upon punched cards were especially concerned with this problem because of the limited amount of storage space on a card. The name of the function correspondingly expanded to information storage and retrieval. (U)

The term information retrieval has now become generic. Some of the most trivial computer processes are referenced as information retrieval processes. In essence, the name implies that data are stored and retrieved by a fixed set of computer instructions. The difficulty is to establish precise limits for a range of different problems, each with an increasing degree of complexity, described by the label of information retrieval. The effect of the generalization of the term is to vitiate its meaning; indiscriminate use dissembles between the original concept applied to the field of documentation and the generic concept applied to any automated procedure for storing and retrieving data. (U)

Even in its broadest sense, however, information retrieval is a misnomer. An automated retrieval system is generally defined as: A system for structuring the information in a collection of documents so as to facilitate the storage and retrieval of information. The difficulty with this definition is that no information, at least in any formal sense of the word, is either stored or retrieved. The focus remains upon classification schemes, and it is classification data that are stored and retrieved. The actual information stays in the collection of documents (except in systems such as Minicard), and the retrieved data simply acts as a pointer to the actual location of the document and its information. (U)

The objective of a Fact Correlation System is to retrieve information. This objective requires a more rigorous definition than the current concepts of information retrieval, which generally limit the problem to the retrieval of information about a document rather than the information within a document. In one sense, therefore, the system may be defined as a technique to store documents and to retrieve documents or specific information within the documents automatically. In another sense, the system may be defined as a technique to analyze information within documents, to isolate unique information and eliminate redundant information, to correlate the unique information with the existing corpus, and to structure the information so as to facilitate the retrieval of either explicit or implicit information in response to a specific request. This definition still lacks rigor; the concept of information is unclarified. The important point is that the information about an event or thing from a set of documents is the primary factor to be considered. (U)

The conceptual framework of this study is the second definition. It is concerned with actual storage and retrieval of information. (U)

B. Analysis of Present Concepts

There has been a considerable amount of research expended upon the problem of information retrieval, particularly as it pertains to documentation. The amount of literature--speculative, theoretical, and practical--is extensive. This discussion presents a cursory review of the concepts of information retrieval. (U)

The problem in information retrieval, whether the subject area is such routine documentation as personnel files or the special requirements of intelligence analysis, is a function of the large volume of available information pertaining to any subject field. Since the advent of the digital computer, it has been hoped that the machine's power to perform logical and computational service rapidly and unerringly could be harnessed to the information retrieval task. (U)

For routine documentation searching, the task has always been essentially clerical. A librarian can be useful in retrieving information, given an effective cataloging and indexing scheme, without understanding the contents of the documents being searched. Thus, for the retrieval of documents originally prepared with an orientation to the user's requirements, it is possible to substitute a computer without sophisticated, semantic or inferential capabilities to perform many of the librarian's functions. At this point the question of effectiveness is not germane. (U)

Present systems of information retrieval are based upon an indexing system in which a document is described by a list of key words or descriptors. (The term, descriptor, is one of many similar concepts; it is used in a general sense in this discussion.) The list of descriptors for a single document is unique, although a particular descriptor may be applicable to many other documents. The document itself is identified by a unique number that serves to locate the document in its storage place. There are two ways in which information about documents may be retrieved--by using the look-up principle or the search principle. The name of the principle may vary among systems. There is no third principle; at best, a sophisticated technique may combine the two basic principles. (U)

In a look-up system document numbers are listed under descriptor headings. Information is selected or retrieved by referring to the record for each descriptor stated in the query and comparing these descriptor records for matching document numbers. A document is selected if it has been indexed under each of the descriptors in the query. A new document is added to the file by placing its number on each appropriate descriptor record. The descriptor records must be stored in known locations within the file and the document numbers must be arranged in numerical sequence on the descriptor records. (U)

In a search system, descriptors are listed under document number headings. Information is retrieved by scanning the entire file (or major blocks of the file). Descriptors of the query are compared against the descriptors on each document record scanned. When a document record matches each of the descriptors in the query, the document is selected.

A new document is added simply by adding a single docum..nt record to the file. Records may be searched in any sequence, and the files need not be arranged in numerical sequence. The search technique is more efficient if descriptors are listed alphabetically on the document record. (U)

Each of these methods has its advantages and disadvantages, and the choice of a method depends upon the intended application. For example, greater retrieval speed may be a determining factor for selecting the look-up system; however, if the files require frequent updating and additions, the search system is the better choice. (U)

A simple information retrieval system has three distinct stages:

(a) Input--the indexing procedure.

(b) Storage--the medium used to store the index record and the methods used to create the record.

(c) Output--the answers to the information retrieval questions.

All or any part of these stages may be processed on a computer. The amount of manual processing involved is inversely proportional to the efficient and accuracy of the system. The storage and output stages are handled fairly well by the present information retrieval systems for simple document retrieval problems insofar as they are completely automated. The input stage still relies heavily upon manual processes in preparing data, keypunching, and verifying. These stages and the fundamental principles of their application are comparable for manual, mechanical, and electronic system. (U)

This review is really not so cursory. Despite the number of different retrieval systems and the extensive literature describing them, the differences are in degree rather than in kind. The processes are essentially the same; only the techniques, the application of principles, differ. (U)

In essence, there are four factors that govern the design of an information retrieval system. Although most of these factors have been discussed generally, briefly stated they are:

(a) Terms.

(b) Structure.

(c) Interrogation.

(d) Evaluation.

There are a number of designations for terms; descriptor is only one. The difference in terms are minor. Their impact is primarily oriented to classification schemes, but processing techniques may be complicated by the requirements imposed by the definition or the magnitude of the term. (U)

Structure refers to the way information is stored; it is characterized by the look-up and search methods. Both are a little more than ordered lists. Interrogation techniques are based upon Boolean functions, although these techniques have also been described in set theory and symbolic logic. The methods of interrogation are equally applicable to manual and automated systems. Manual systems allow many Boolean functions to be performed at the discretion of the user. But most automated systems that have been actually implemented are limited in their interrogation to <u>and</u> functions. (U)

Evaluation functions are an orphan. Several systems have been thoroughly tested and evaluated, usually unfavorably. But no known system includes self-evaluation procedures. (U)

Information retrieval appears to be characterized by only a few basic factors. If this hypothesis is valid, it is also apparent that none of the factors has been probed extensively. Nor have the interrelationships among the factors been fully analyzed and exploited. The extension of information retrieval in any sense depends upon the development of significant new concepts within the framework of these factors. (U)

C. Conceptual Basis of Fact Correlation

Information is conveyed by a variety of methods in human societies. But the most significant method is the medium of words, the symbols that express the thoughts of individual human beings. Facts are a form of information. They occur in diverse contexts, often becoming significant only after a series of facts have been correlated. But the essential feature is that the exchange of information depends upon a set of processes performed upon the set of words that convey the information. These processes are performed by both the sender and the receiver of information. (U)

Words and facts, which are generally an ordered set of word, are conveyed through language. A language can be learned; so can the information expressed by the language. Thus, there are two rudimentary problems associated with the correlation of facts--language and learning. (U)

Neither the language analysis nor the learning theory necessary for a man-machine system has been sufficiently studied and formalized. It is, however, possible to conceive of a man-machine system in which an automata is organized to analyze language to derive the information from a set of statements and to perform learning functions to correlate the information. The degree of sophistication attained by the automata depends upon its interactions with man and its environment. (U)

Between the statement of the concept and its formulation, there is a significant gap. Initial research indicates that it may be feasible. The remainder of this report discusses the problems implied within the conceptual framework and the degree of research necessary to achieve the objective of developing a Fact Correlation System. (U)

III. SYSTEM ANALYSIS

A. Objectives of System Analysis

The objective of system analysis is to ensure an integrated approach to the design and development of a system that will satisfy U. S. Air Force requirements for the correlation of facts, such as those encountered in intelligence data; at the same time, such a system must be a highly effective and efficient operational system. To fulfill this objective requires the following types of activities:

(a) Determination of user requirements.

(b) Determination of system requirements. (The formulation of (a) and (b) together constitute the system concept.)

(c) Determination of system characteristics, capabilities, and constraints.

(d) Development of a functional description of the system and the interactions among equipment, computer programs, and personnel.

(e) A statement of design and development requirements.

(f) Development of the relationship between system tasks in order to monitor the design work to ensure fulfillment of user requirements within the framework of an effective operational system.

(g) Coordination and supervision of the development of hardware specifications, computer programs, and system utilization procedures to ensure system operational effectiveness. (U)

System analysis provides the framework for research and development efforts in the several tasks or functional areas that comprise the system. Adopting and maintaining the system viewpoint permits the development of over-all system requirements and their implementation without considering the details that are necessarily encountered by the individual researcher in performing his task. Maintaining a system viewpoint of over-all requirements will ensure a properly integrated set of system functions or tasks, constituting an effective operational system. (U)

The system analysis task also serves as a means of technical project control and system management of the research and development effort. One of the most important aspects of the coordination of the research and development is the proper long-range planning and phasing of tasks. This function promotes the development and timely completion of intermediate products or subsystems that have immediate applicability and can be utilized without awaiting development of other portions of the ultimate system. As these intermediate products are completed, it may be possible to combine them without undue extra effort, so that the system can be construed as a sequence of building blocks, each providing significant added capability as it is completed and incorporated in the system. The other method of improving capability by further developments and refinements within the building blocks can also be utilized wherever it is fruitful. (U)

The remainder of Part III discusses the activities required to fulfill the objectives of system analysis and presents a purview of the system as envisioned at this stage of analysis. (U)

B. Methodology for System Analysis and Design

This section briefly discusses the methodology to be used in fulfilling the objectives of system analysis. A general methodology for the systems analysis task has been outlined in the statement of objectives. A more detailed picture of the methodology to be used in system planning and analysis and in project and task control is illustrated in Figure 3-1. The

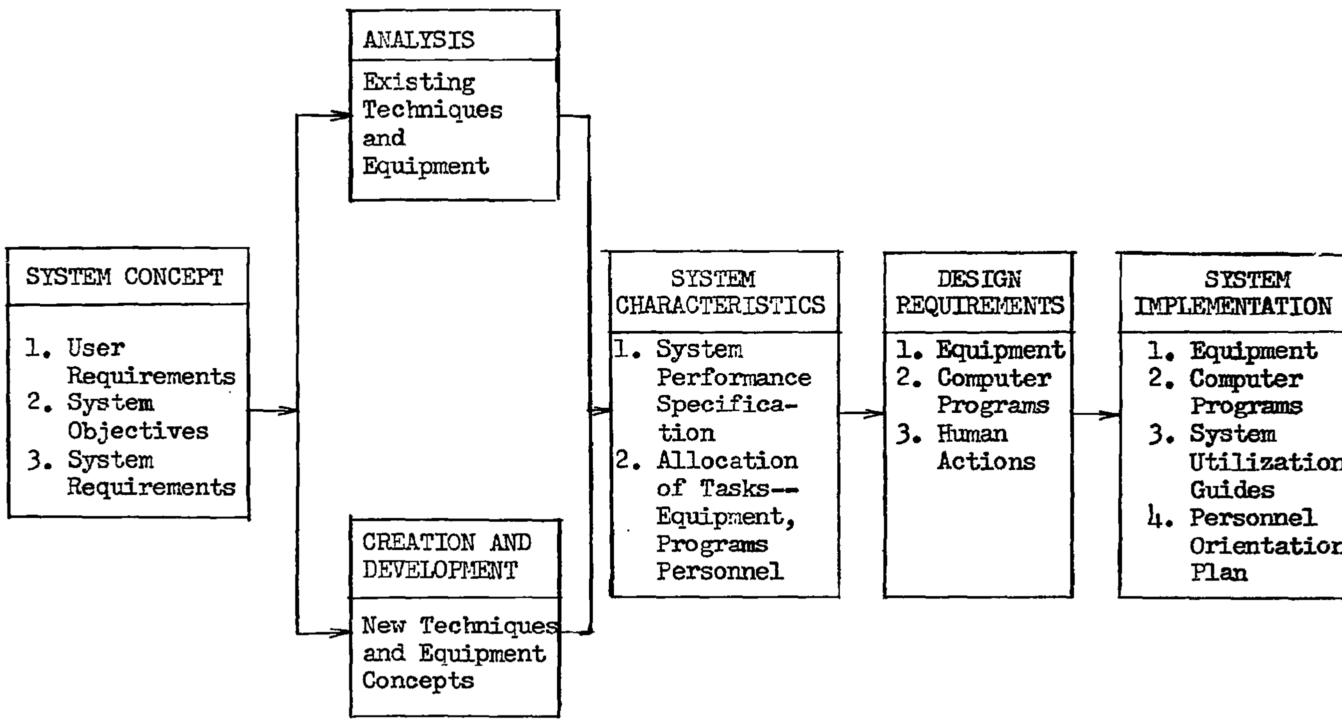

FIGURE 3-1. Phases of System Design and Development

work is divided into six phases of activity. The over-all methodology emphasizes system analysis and design as distinct from the methods to be used in performing the tasks within each phase. The latter methodology is a product of the professional skills and background of the individual personnel who contribute their knowledge to the analysis, design, and development of this system. (U)

The development of a satisfactory operational Fact Correlation System depends upon the first phase, the development of a suitable system concept. Unless the system objectives, user requirements, and derived system requirements are correctly formulated and effectively stated, further development of the system will be severely hindered. A workable system could perhaps be developed, but it would not do the job that the Air Force expects. Coordination with the proper commands or agencies and their personnel is essential, if not vital, for this phase of system activities. (U)

Once the system concept has been formulated and stated, the means of implementing it must be sought. This activity involves the analysis of existing techniques and equipment and the creation and development of new techniques and equipment concepts. Various continuing studies have indicated generally that existing equipment for performing the functions likely to be necessary for a Fact Correlation System is adequate, except possibly for input-output devices. Conversely, existing techniques for language analysis and synthesis, adaptive learning, and fact correlation are inadequate. Consequently, the creation and formulation of new techniques constitute the primary research problem in developing an effective Fact Correlation System. (U)

An important aspect of this research is to recognize and delineate the numerous, diverse, and complex tasks that will be performed by such a system. First, the major system tasks or functions are delineated. This step is followed by finer subdivisions of system functions into subtasks until some reasonably basic atomic level is reached. This functional analysis provides a sense of direction to individual research efforts and aids in ensuring consideration of all features necessary for an operational Fact Correlation System. System analysis fulfills its integration function by viewing the tasks as closed entities or "black boxes" and confines itself to specifying inputs and outputs for each task as well as defining the interrelationships and feedback requirements among all tasks. If the task breakdown is effective and input-output and feedback requirements are met for all tasks in the system, then the operational performance of the system is assured by the system analyst without consideration of the details of design within tasks. (U)

A comparison of current research with the functions and tasks required by an adequate Fact Correlation System, even at a relatively general level, reveals a lack of awareness of some of the problems involved. An awareness of the nature and scope of some problems is either nonexistent or vague, and the necessary research in these areas has just begun or has only been outlined in a general way. What is urgently required to improve this situation is to redefine the problem of data correlation and retrieval in terms sufficiently broad to include all the levels of language analysis that are necessary and to stress the adaptive and dynamic aspects of the system. (U)

From the analysis of existing techniques, their further development and modification, and the creation of new techniques for language analysis and adaptive fact correlation comes the formulation and definition of the system characteristics. The system characteristics are functional descriptions that distinguish system capabilities and limitations and constitute a system performance specification. The delineation of the system characteristics includes the allocation of tasks among personnel, computer programs, and equipment. (U)

In the design requirements phase, detailed specifications are developed for operational equipment requirements, computer program requirements (operational and executive), and human action requirements. There are two types of specifications. The first set is performance specifications, which contain the design requirements (or what the system must do in terms of equipment, programs, and human action requirements) at the most detailed level. The second set is design specifications, which specify the detailed procedures for meeting the performance specifications in terms of equipment design specifications, program design specifications, and detailed operational procedures. (U)

The final phase in the development of an operational system is to implement the detailed specifications by developing operational and executive computer programs, a personnel orientation plan, system utilization guides (which contain effective methods and procedures for system operation and system evaluation), and the procurement of necessary new equipment. (U)

C. Assumptions, Environmental Constraints, and Limitations

To provide a framework for the continuing research and development efforts of system designers, a set of environmental and system assumptions must be formulated and stated. (U)

The system assumptions and environmental constraints cannot be completely specified until a thorough analysis of the user requirements has been completed. Assumptions and constraints may also be modified by the results of research and development efforts which alter potential system capabilities. (U)

The following assumptions have been formulated as a basis for the system and task analyses of this report:

(a) High level or complex interpretation of either raw or correlated data will not be a computer function, but a human function.

(b) Inputs of the Fact Correlation System will be limited to English text and numerical data. Recognition and interpretation of pictorial and graphic data such as aerial photographs and maps will not be a system requirement, except insofar as these data are inserted in linguistic or numerical form.

(c) Each instance of entry into the system of input information concerning a given topic or area will be limited to approximately 2000 words of English text. There will be no constraint, however, on the number of sequential instances of entry from either one or several documents.

(d) A primary system requirement will be continual man-machine interaction for purposes of validation, clarification, interrogation, file searching, and adaptive learning by both men and machines.

(e) The essence of this system involves the retention, correlation, and retrieval of factual data on the basis of related context and meaning, rather than any documentary or literary source relationship.

(f) The system will not begin operation with a large corpus of background information and fixed operating, processing, search, and correlation techniques. The memory structure, search, and correlation techniques of the computer will be adaptive and the evolving corpus of correlated information will be dynamic. That is, the system capabilities will be developed principally through the process of adaptive learning applied to a growing corpus of facts. (U)

The first assumption reflects the fact that an extensive knowledge of diverse aspects of the human situation in the world is necessary for the proper evaluation and high-level interpretation of intelligence data. This condition is particularly evident when political, economic, and psychological factors pertaining to a situation are considered. In addition, a large amount of expert knowledge in specialized fields is required. Only the human mind can satisfactorily perform such interpretations at present. The objective of an automated Fact Correlation System is primarily to extend the effectiveness and scope of the efforts of human analysts. It is worthwhile to emphasize, however, that the automated system may be capable of quite sophisticated correlations as distinct from interpretations. It is also anticipated that the degree of knowledge or expertness of the system in fields related to the inputs it receives will constantly increase. However, the sophistication of its knowledge and associating and relating structure cannot approach that of human beings. (U)

The second assumption means that the system will not be provided with the capability to recognize map structures or other geometrical patterns or configurations. It will, however, be able to process and correlate information that is contained in the relationships among symbols. The importance of photographic interpretation is fully recognized. It is also recognized that a computer may be used to interpret photographs. The assumption was not made because of a failure to realize either of these facts, but because this capability is considered to be beyond the scope of this report. (U)

The third assumption establishes a tentative limit on the number of input sentences from a given document on a given general subject that the system will analyze and incorporate into its internal corpus in a single computer run. Any number of these entries can be made sequentially with interruptions for processing. (U)

The fourth assumption is an assertion that a dynamic system, which is capable of adaptive learning and of answering more or less unrestricted queries, must rely on interaction with human beings to assure effective performance. (U)

The fifth assumption emphasizes that the system will be designed to correlate and retrieve the related content of statements. It will not be concerned with the selection of related documents, titles, or abstracts as in conventional systems for the library-type retrieval of indexed information. The "information" that is correlated and retrieved must be measured in pragmatic terms; namely, its value or utility to the user. The information conveyed by the system must be defined in terms of the increased knowledge and effectiveness gained by the user. The term "information" is not used in the sense of C. E. Shannon;[1] that is, in terms of statistical rarity of messages. (U)

---

[1] Shannon, C. E., and Weaver, W., Mathematical Theory of Communication.

The last assumption states that the system will not be given a voluminous body of initial information of the type that a human expert in some activity might have at his disposal. However, the system will be influenced in its mode of development by the particular data that it receives as input. (U)

D. User Requirements

The first task in the design and development of a system for the correlation of factual data is a comprehensive analysis of present and projected user requirements.(2) User is a generic term meaning, in this instance, those personnel who ultimately receive and use system outputs. It does not mean an operator who enters inputs and actually "runs" the system. Of course, a system user can also function as a system operator. Such an analysis is called an operational analysis and results in the development of operational requirements. As distinct from this type of analysis, system analysis is concerned with the development of system requirements, which are general statements concerning ways to fulfill the operational requirements. (U)

An effective investigation of user requirements may entail an operational analysis of the anticipated application (e.g., the process of analysis of intelligence data). The ultimate scope of this operational analysis will be determined by the extent to which the present techniques for intelligence data analysis and correlation are applicable or adaptable to

---

(2)See also: Herner, Saul, The Relationship of Information-Use Studies and the Design of Information Storage and Retrieval Systems.

an automated Fact Correlation System. Although an automated system should be designed to complement the user, such a system should not be constrained by the simple adaptation of existing manual procedures, which are possibly outmoded or inadequate. (U)

A serious problem of information retrieval faces any potential user of facts and data that are located in some unknown portion of published corpus. The inadequacy of the traditional methods of information retrieval, based upon the retrieval of documents rather than their factual content, is well known. For routine documentation searches, the task has always been essentially clerical. A human librarian can be useful in the retrieval of information, given an effective cataloging and indexing scheme, without understanding the contents of the documents being searched. Thus, for the retrieval of documents originally prepared with an orientation to the user's requirements, it may be possible to substitute a computer without sophisticated language analysis or inferential capability to perform many of the librarian's functions. (U)

The intelligence analyst, on the other hand, cannot conceivably perform his function of gleaning unemphasized, possibly implicit, information from documents without completely comprehending their contents. In order to provide information (as opposed to documentation) in response to an intelligence request, a computer system must, therefore, have a high order of semantic and inferential sophistication. Since the purpose of an intelligence system is to extend the performance and effectiveness of individual human analysts, such a system must be the goal of this project. (U)

The scope of information retrieval can be considered as a spectrum of retrieval problems, as shown in Figure 3-2. If the specific problem or subject field occurs at the low end of the spectrum (left side), the present information retrieval systems are adequate. Problems of the library retrieval type have been successfully handled by data processing systems, although there is still much room for improvement in developing more efficient retrieval systems with greater storage capacities. As the problem area or subject field approaches the high end of the spectrum (right side), the inadequacy of the state-of-the-art in information retrieval and the need for research into new techniques becomes more apparent. An information retrieval system capable of deriving, correlating, and retrieving explicit and implicit factual relationships is envisioned as the ultimate goal for the Fact Correlation System. The requirements of the user are the primary consideration in determining what type of information retrieval system is needed: the more he recedes from the high end of the spectrum in establishing his requirements, the less sophisticated the system needed. (U)

One difficulty in an intelligence analysis system comprised of human beings is that of communication. Because of the structure of a human organization and the limitations of human nature, it is quite probable that an intelligence expert will receive only a limited amount of data, judged by someone to be appropriate to his field of competence and his special assignment in intelligence work. One of the advantages of an automated intelligence system is that any human analyst associated directly or even indirectly with the system will have immediate access to all inputs entered

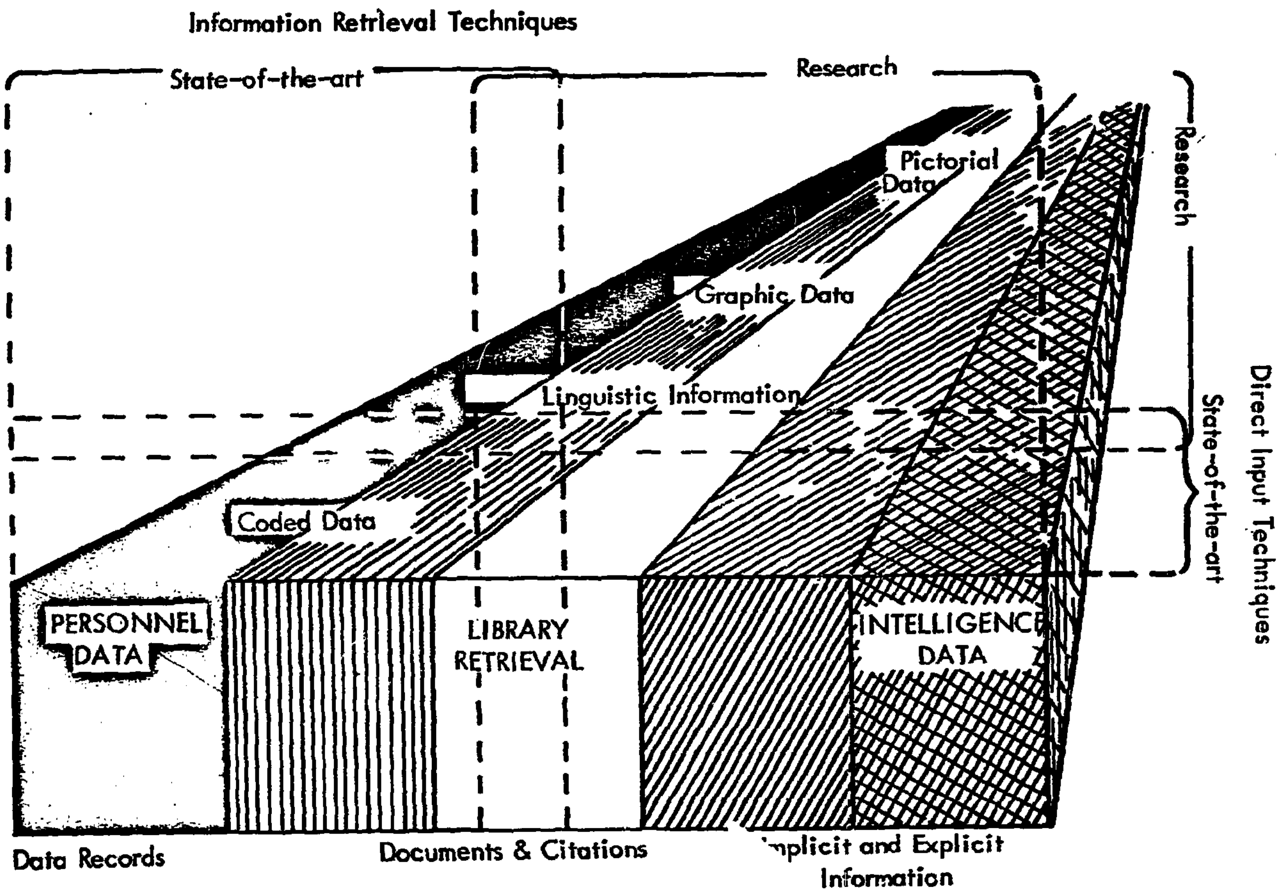

FIGURE 3-2. The Spectrum of Retrieval Problems

in the system (subject to security restraints). This fact alone will facilitate the human analyst's task considerably. (U)

Fundamental to the development of such a system is the interaction between man and computer. The operational analysis in this project will involve an intensive study of the nature, form, and volume of data to be handled by the system. Particular attention should be paid to the nature of user requests for information, to the logic of questioning, and to the human guidance that can be provided to the computer. A retrieval system with relatively few elements in the initial corpus will contribute to the development of a specialized data base for the user. A query translator, which is sufficiently sophisticated to minimize linguistic constraints on the structure or format of queries or interrogations, will simplify the process of interaction. The major tasks in the development of user requirements are:

(a) To analyze present procedures, concepts, and theories related to the analysis and correlation of factual data.

(b) To ascertain and evaluate the applicability, adaptability, and status of these procedures and concepts.

(c) To select those concepts and theories, as well as practical solutions, that contribute to the unified development of an automated Fact Correlation System.

(d) To establish new procedures specifically adapted to an automated Fact Correlation System.

(e) To establish guidelines for continuing research into projected user requirements. (U)

E. System Concept

The ultimate objective of this program is to develop automatic techniques for the processing of textual data; the processing function includes both the storage and the retrieval of information. This objective requires the development of a new conceptual basis for the retrieval of information, a basis that circumscribes the limitations of the current state-of-the-art in information retrieval. (U)

The basic frame of reference of this study is the function of intelligence analysis. An expert in intelligence has a goal--to discern the implications of random sets of information in order to discover their essential characteristics. Once the information has been correlated, the analyst derives conclusions about capabilities, situations, purposes, and intentions. In performing these functions, the analyst uses a special methodology. (U)

Since an ancillary purpose of this program is to facilitate the development of a generalized system for correlating facts, it should be adapted to the stringent requirements of analyzing intelligence data. A system capable of correlating facts under the conditions that apply to intelligence data, namely random and sometimes confusing and contradictory data on diverse subjects from diverse sources, will have the capabilit· of correlating facts from a more orderly environment. Nevertheless, the development of a particular system will be affected strongly by the application for which it is used. Therefore, criteria for correlations should be developed by man-machine interactions in terms of the particular input data and system environment. (S)

A basic tenet of this report is that the "documents" that comprise the corpus of data will be quite numerous. Each document will contain a high density of information about a particular set of events or circumstances. The information in each document must be correlated with information that has been collected and stored within the memory or files of a computer. However, the utility of data has a certain decay rate as a function of time in storage. The system must provide some measure of the utility of information as a function of frequency of use. This measure will be useful in saving computer storage space by eliminating obsolete or inapplicable data. (U)

These assumptions presuppose that the relevant information is not known _a priori_, an implication that the correlation of facts within the corpus of data is a dynamic operation. Establishing a link between new bits of information and the corpus will be difficult unless the system allows a computer to request further information whenever a gap in information occurs. Therefore, the process should also be capable of interrogating a human analyst whenever specific data is required to facilitate the process of assimilation. (U)

Once information has been stored in the form of basic symbolic representations--that is, words--subsequent recognition of similar symbols is a routine operation. The treatment of phrases is more difficult, but as soon as definite patterns have been established, some groups of words could be recognized as definite phrases. The recognition of sentences is another problem entirely, since the probability is low that even the same bits of information are generally expressed in exactly the same sentence form. It is highly likely, therefore, that the recognition of sentences will defer to the recognition of similar thoughts expressed in different ways. These thoughts must then be equated by storing them in a single unique form. (U)

The explicit form of any sentence is usually unique. The relationships implied by an explicit statement depend upo.. the assimilation of an ever widening set of relationships. The scope of this problem is broad; its solution depends more upon the application of techniques of learning and automata theory than upon the familiar concepts of information retrieval. It is in the area of correlating explicit and implicit meanings that the key aspects of the problem lie. (U)

These concepts indicate that an adaptive or self-organizing system should be developed in order to correlate new information to the corpus of data. The process of correlation or assimilation corresponds directly to a rudimentary ability to learn the information. In short, learning is the crux of adaptation. These requirements do not imply that a self-organizing system must be able to solve problems; the primary criteria for correlating facts is the ability to isolate new items of information and to assimilate them into the corpus. The corpus may then be modified on the basis of significant new information. An orderly process of learning again suggests that interaction with a human analyst will enable the information system to assimilate its information more readily. (U)

The basic system concept is illustrated in Figure 3-3. Although detailed equipment and techniques required to implement it will be discussed later, a general purview of the system concept follows. (U)

The ultimate concern in designing a system is the effect upon the

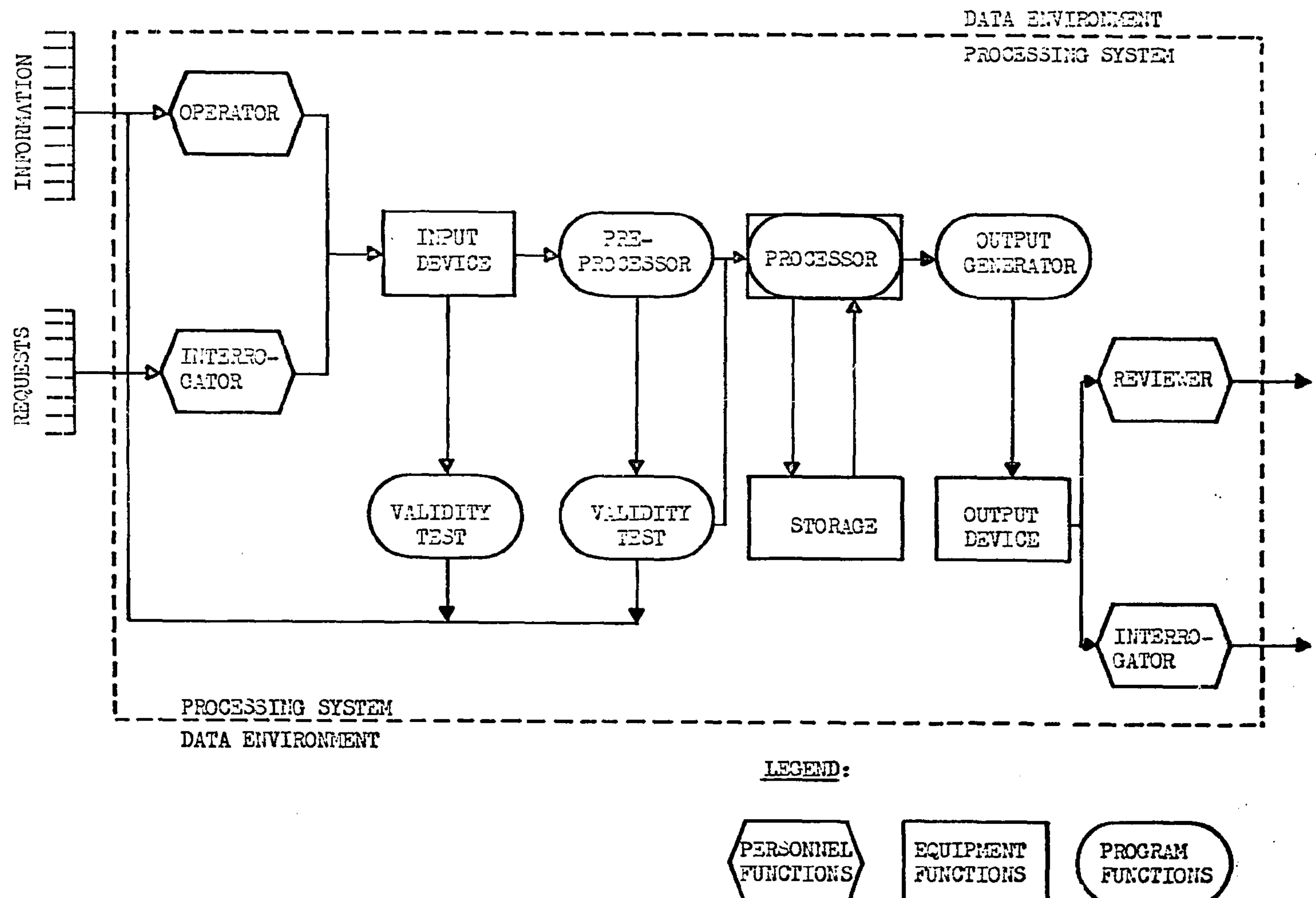

FIGURE 3-3. Basic System Concept

user, the personnel who submit information to the system or request information from the corpus. If the user is an analyst familiar with the development and general contents of the system, then some restraints may be placed upon this user. However, if the user is generally unfamiliar with either the contents of the system or with its operations, then the type and number of restraints should be reduced to a minimum. (U)

As a user of the system, the analyst's task will include selecting information to be entered into the system, acting as the agent to interpret questions in the system, and requesting information from the files. When the analyst acts as the "teacher" of the system, responding to questions pertaining to the correlation of facts, then the questions posed to the analyst should be framed so that the intent of the question is clear. These considerations indicate the following factors:

(a) That the system print out the ambiguous information in its original form.

(b) That the system ask a particular question about the information.

(c) That the reasons for the questions be listed.

(d) That the possible interpretations on the basis of past information be listed.

With this data, the analyst should be able to recognize the problem of correlation confronting the system and to respond with an acceptable interpretation. Insofar as possible, the responses from the analyst should conform to one of the possible interpretations reached by the system. If this response is impossible, then a new interpretation must be clearly stated by the analyst. (U)

The selection of information is a function of the particular application of the system. Since information is gathered from various sources and in various forms, it may be necessary to transcribe the information into a more manageable format. Optimally, the raw information should require no human intervention after it has been collected. At present, processing functions such as editing and keypunching constitute both a bottleneck and an expense. If these functions can be reduced or eliminated, the efficiency of the system will be enhanced. The obvious solution is a reading device that can directly convert textual data into digital form. (U)

Once the data is entered into the computer, a series of preprocessing functions should validate the input data. This function includes such basic tests as spelling, typographical omissions, inadequate character recognition, and punctuation. On a higher level, questions of morphology and word order should be incorporated. Whenever discrepancies arise and self-correcting programs fail, the system should request correct information from an analyst. (U)

Mechanical errors can be resolved by a preprocessing stage. Once information enters the correlation stage, which includes a series of processes performed sequentially, the feedback loop is no longer concerned with mechanical errors. At this time, the learning processes should interrogate an analyst to eliminate apparent ambiguities that have arisen during the analytic process. For purposes of later fact retrieval, multiple renderings of ambiguous texts could rapidly lead to an unacceptably large number of logical branchings in analyzing the explicit and implicit information content of intelligence data. To the extent that automatic ambiguity reduction is lacking, it is necessary to include provisions for interaction between human analysts and the computer--in order to aid the computer in reducing the ambiguity that remains in humanly generated texts after automatic ambiguity reduction techniques have been applied. (U)

The qualification for the necessity of human intervention in an "automatic" fact correlation process is not a rejection of the basic premise that study directed to such a goal is meaningful. Current research on techniques of automatic semantic ambiguity reduction represents a substantial advance over what is possible with purely lexical and syntactic methods.[3] The human element is included in the basic system concept because it is currently inconceivable that perfect ambiguity resolution can be preprogrammed on the basis of *a priori* understanding of human linguistic behavior. (U)

A second type of feedback loop consists of a verification process. When the corpus is small, this process plays an important role in assessing the correlated data. As the corpus becomes larger, the role of the verification function will be reduced as the learning process establishes more correlations between various bits of data. (U)

The processor is, of course, the heart of the system. It is the vehicle for structuring memory, performing adaptive learning, incorporating new data, correlating facts, and retrieving data in response to human

---

[3] This approach is more fully described in Sommers, F. T., *Semantic Structures and the Automatic Clarification of Linguistic Ambiguity*.

interrogations. Usually any consideration of computer operations emphasizes the speed of modern computers. The primary emphasis in this program will be to determine the most efficient computer speed as a function of system requirements. Under normal conditions, it is possible to assume that information will be received by the system at a steady but slow rate and that queries will occur randomly in time. If the executive function of the computer programs can interrupt normal information processing to respond to a request, it appears that the high computer speeds are not an essential prerequisite. (U)

Another factor to be considered in the system is storage capacity. This consideration is a function of the amount of detailed information that will be retained in the corpus with the passage of time. System capability will, of course, be affected by the types of storage devices used. The capability will be determined by a trade-off between cost and speed as reflected in access time. A more important factor is the efficient use of storage. Efficiency can be achieved by the development of structural data classification schemes that include the important factor of frequency of use of the data. This frequency factor would be modified continually by the system as a function of use over a period of time. A large class of statements provides significant, useful information only when specific dates, times, and geographical locations are known for events described by the data. This condition is significant in the case of intelligence information. (U)

The sentence generator performs the language synthesis necessary to convert information retrieved from storage by the processor to meaningful, factual statements in natural language. This statement formation function is the converse of the language analysis function performed by the pre-processor. The sentence generator will ensure that output statements are characterized by the same types of validity as the pre-processor checks in its analysis. (U)

A final consideration is output equipment. The essential requirements are primarily related to legibility and efficiency. Printer speed is not a serious problem since most of the system output will be relatively short. The output device completes the linking of the system to its environment. The human analysts perform the function of reviewing, interpreting, and evaluating system responses, preparatory to the input of additional data or further queries into the system. (U)

F. Equipment Functions

The Fact Correlation System will be comprised of human analysts, computer programs, and equipment. Human beings are the most flexible and intelligent of these components, while equipment is the least flexible and intelligent. Conversely, equipment performance is the most reliable and most "automatic," while human performance is generally the least reliable or accurate. One of the major assumptions in this approach to the design of the system is that the "intelligence" of the computer programs, with human interaction, can be increased substantially. However, equipment flexibility remains highly restricted. Increases in its flexibility beyond a certain point are obtained only at high costs in money, development time, maintainability, operability, and reliability. These factors determine, in a general way, the allocation of tasks among equipment, programs, and human beings. (U)

Before discussing the specific equipment functions, it is pertinent to mention that the general flexibility and utility of the system can be extended by the use of remote input-output sites. For various reasons, the basic processor, control, and storage equipment require a central facility. However, remote input-output equipment such as request panels for computer interrogations and small display panels for recording system responses could be located in the offices or work areas of the human analysts. This configuration would facilitate immediate man-machine interactions such as interrogations, higher-level error checking, and evaluation of system responses. Advances in input-output equipment may make such a system configuration more economical in the next few years. (U)

The system functions that are relegated primarily to equipment are noted in Figure 3-3. The first equipment component required is an input device. One of the principal problems in existing data processing systems is the relatively inefficient speed of input-output devices. This problem is further aggravated by the input requirements for linguistic information. Present methods for data input result in serious losses in time and money. The efficiency of a Fact Correlation System will be enhanced by the incorporation of an automatic reading device that converts printed text into digital form. (U)

The next major equipment functions are performed by the processor and the storage. The critical elements of this system in the processing and storage areas are associated with techniques rather than hardware.

The processing speeds of existing electronic data processing systems are quite adequate to handle the volume of data anticipated. Ultimate storage requirements will be determined by the scope and magnitude of the data to be included in the system. Together with the relatively large size anticipated for the computer programs, storage requirements may be large; but this amount of capacity is not expected to be a major problem. The state-of-the-art of processing and storage equipment is sufficiently in advance of language analysis, learning, and programming techniques that a radical advance in the capability of this equipment will not be critical at least until break-throughs in techniques have been achieved. (U)

The last equipment function in the system is that performed by the output devices that transform digital data inputs from the sentence generator into sentences in natural language. This process will not involve a high-volume or high-speed operation. Nevertheless, in common with the input device, advancements in the equipment state-of-the-art may be necessary to perform this function adequately. (U)

The equipment in the system is the carrier of data and the means of implementing the system intelligence in an operational sense. It is also the memory of the system. It is clear that the automatic operation and capability of the hardware internal to the system presents no problem. The critical equipment functions are those involving man-machine interactions. Considerable effort must be devoted to the study and resolution of the automatic input of data. Even more important is the assurance of having suitable equipment available for the rapid input of human queries and interrogations, which must reach the control element of the computer quickly in order to elicit relatively prompt responses. The equipment must also have the capability to accept and route isolated and relatively short additions to its corpus. These additions will rectify or clarify ambiguities or gaps in system knowledge that prevent further system operation or growth at any particular time. (U)

G. Program Functions

The system concept in Figure 3-3 shows indirectly the computer program functions. Those parts of the system labeled with a technique symbol represent tasks requiring the development of analytical techniques that will be converted into computer program functions. The entry of inputs into the system by the human operator is both a personnel function and a program function. The program function here is not complicated. It requires only that the appropriate rules be followed for entering data on punched cards or magnetic tape. If, as is more likely in this system, an automatic text reader is used as an input device, then there is virtually no program function. The appropriate rules for entering data are a human function constrained primarily by the capability of the text reader to recognize certain standard typewritten or printed symbols. (U)

On the other hand, the interrogation of the computer by the human analyst is a major program function. Questions will have to be analyzed linguistically and understood by the computer so that the appropriate associated classes and items of information can be retrieved and meaningful responses generated. The more versatile these programs are, the fewer will be the format and form restraints on human queries. If the question analysis and retrieval programs lack generality and versatility, then the restraints on the human interrogator will be greater and will force a more rigid and restricted form on human queries. (U)

The validity testing of inputs is a program function. It will test for random mechanical human errors in input preparation as well as machine reading errors. In addition, it must check for errors in spelling, grammar, sense, and logical consistency. It must incorporate automatic procedures for communicating the exact nature of the input errors to the human environment for corrective purposes. (U)

The pre-processor is primarily concerned with the linguistic analysis of inputs from spelling to grammar to sense significance to logical consistency. These tasks are entirely a program function. In addition, computer programs must extract, retain, and store the factual content of the inputs. (U)

The processor is another major program function. It must be capable of intelligent behavior, in the sense of adaptive learning, storage, and response. It must be capable of accepting converted symbolic facts from the pre-processor and incorporating them into a general stored fact structure representing the state of knowledge of the computer. This function requires a dynamic, adaptive capability to discriminate, associate, relate, and generalize. (U)

Procedures for searching for and retrieving information from storage are program functions. This function is vital for the intelligent response to human queries. The processor together with the techniques for computer memorization of symbolic data and relationships, and techniques for the appropriate selection, association, and retrieval of data for responding to specific queries is the heart of the computer program functions and the entire system. (U)

Ultimately, learning depends upon the ability to organize and to memorize information. An efficient memory structure improves the recall and juxtaposition of information in order to verify past experience or extrapolate new associations. This fact emphasizes the need for an adequate memory structure. (U)

Computer programming has placed a major emphasis on the organization of data. However, most of this work is based upon a different set of requirements from those of fact correlatio. . For example, an alphabetical list of words can be scanned rapidly to locate a particular word, and large groups of data can be numbered so that the data can be indirectly referenced. But these examples indicate the limits of such structures, which are their failure to allow for interrelations among separate items in memory. (U)

Criteria for memory structure should be established such that the selection of one word automatically excludes any unrelated words and includes all related words. Programming techniques alone will not satisfy these conditions unless an adequate theory is devised for structuring memory. (U)

The next computer function is the sentence generator, which consists of mechanical programmed procedures for converting machine responses to valid English sentences. This function is an inverse of the pre-processor.

It is a set of synthesis rules (whereas the pre-processor is an analysis function) for constructing of valid English summary sentences. From the retrieved data for the response, it must construct summary sentences that contain correctly spelled English words, and that are syntactically valid, sensible, free of logical contradictions, and factually correct according to the present state of knowledge of the computer. (U)

The review, analysis, and utilization of the output is partially a program function. Computer programs must be developed to provide the human analyst with tools to understand and evaluate system responses and growth, so that its performance can be monitored and evaluated on a continual basis. This function requires a certain set of analysis, diagnostic, and sampling or other evaluative routines to assist human efforts in system evaluation. (U)

H. Personnel Functions

The primary role of the human being in a Fact Correlation System is to provide an effective link between the system and its environment and to improve system capability and performance. The principal functions are:

(a) To select and prepare input data.

(b) To request specific information from the computer.

(c) To interpret and respond to computer queries.

(d) To use and evaluate system outputs. (U)

The selection of input data for the Fact Correlation System is a human task. The complexity of this task depends upon user requirements and the form of data associated with a particular system application. It has been assumed that additions to the factual corpus will be confined to 2000 or fewer English words in one entry. At times this limitation may require abstracting and summarizing from larger documents. This type of activity must fulfill at least two requirements:

(a) It must not be so slow that it degrades system operations.

(b) Abstracts must not omit any essential facts.

The abstracting could be performed by human analysts or by the pre-processor of the computer. Abstracting on a large scale by human beings would probably not satisfy either of these requirements. On the other hand, it is a major problem for the computer to fulfill the second requirement. Nevertheless, the only feasible scheme for a Fact Correlation System is fact summarization by the computer. Hence, the limitation on the size of entry should be regarded as a somewhat arbitrary but convenient unit for system processing. This restriction is partially determined by the requirement for interrupting processing to respond to interrogations entered by an analyst. (U)

The manual preparation of input data is undoubtedly a bottleneck in the operational capability of the system. However, if an automatic text reader is part of the system, data can be entered in the form of ordinary typewritten copy. It is intended that the restrictions on the format of input sentences be minimal. This objective can only be achieved if the language analysis programs are of sufficient power and versatility to handle complicated sentences. (U)

Another key role of system personnel is to interrogate the computer and to respond to computer queries. The first of these roles comprises the ultimate rationale of the system. The second assures validity and effective system growth and operational capability. The objective for the human analysts is the capability to enter unrestricted queries into the system. This objective does not imply that the system will have the capability to answer satisfactorily any and all human questions immediately. Interrogation and communication is a two-way process. The first essential is that the computer be able to interpret unrestricted interrogations to some degree. If the computer is unable immediately to understand the human request for information, in the sense that it can retrieve the relevant information and respond with reasonable effectiveness, it will interrogate the human in order to:

(a) Request that the classes of relevant data be defined with greater precision.

(b) Improve its data search techniques.

(c) Obtain criteria for plausibility with respect to inferential processes.

(d) Inquire about the validity of relationships it has formed.

Another type of computer query occurs when linguistic validity or unresolved ambiguity is in question. A human being must assist in the resolution of such problems. (U)

The last man-machine interaction is concerned with the human utilization and evaluation of system outputs. This interrelationship is the final feedback loop for influencing future inputs in order to improve the computer's state of knowledge and otherwise upgrade system performance. The human analysts will be provided with techniques for refining their interrogations and for requesting processes that will lead to the operation of computer programs that assist in an evaluation of system performance. These programs will present the results of a computer "self-analysis" that summarize its state of knowledge and its present mode of operations. Of course, the human being, by virtue of his superior judgment and knowledge of "content" and environment, will be the ultimate system evaluator. (U)

The system utilization guides must contain operational procedures for system evaluation by the analysts. The development of such procedures is contingent upon the establishment of complete user requirements and system requirements that fulfill them to the greatest possible extent. Only then can expected system performance (quantitative and qualitative) be established. Once it is known what is expected of the developed operational system, performance criteria will be provided and measures of system performance will be developed. (U)

I. System Characteristics and Requirements

This section conveys preliminary views concerning anticipated characteristics and requirements for an effective operational Fact Correlation System. (U)

The correlation of information to extract the essence of meaning in both explicit and implicit relationships has two distinct aspects. The first is the problem of man-machine communication. The second is the development of at least a rudimentary machine intelligence. With current processing techniques, these two aspects of the over-all problem constitute a dilemma. Man could develop more intelligent machines if he could communicate with them better; he could communicate with them better if they were more intelligent. (U)

Communications are conducted through the medium of language; the essence of the communications problems is to extract the significant items of information from a formal language. Generally, despite the amount of work that has been done, the existing work on language analysis remains insufficient to form a basis for computer processing. Thus, this broad area constitutes one of the principal tasks in developing techniques for correlating facts within a body of information. (U)

The problem of machine intelligence is primarily a problem of adaptive learning. Current techniques in learning related to computer processing are primarily limited to problem solving. These techniques are insufficient for the general problem of fact correlation. Thus, this general area is the second major task necessary for the resolution of techniques for correlating information. (U)

Within the framework of man-machine communications and adaptive learning, a number of system characteristics have been envisioned. Six of these were given as assumptions and constraints. Other characteristics and capabilities are given below. H and M indicate human and machine tasks, respectively.

(a) Select input information. (H)

(b) Format inputs, if necessary. (H or M)

(c) Interpret and respond to machine questions. (H)

(d) Interrogate machine by requesting specific information. (H)

(e) Interpret and evaluate machine responses. (H)

(f) Convert English text to digital form. (H and M, or M only)

(g) Validate inputs in terms of spelling, character recognition, punctuation, syntax, sense, and logical consistency. (H and M)

(h) When inputs are invalidated, specify nature of error and request new valid inputs. (M)

(i) Recognize, manipulate, and retain:

(1) Syntax relations (M)
(2) Sense relations (M)
(3) Logical relations (M)
(4) Factual relations (M)

(j) Organize and memorize factual content of input statements. Store in a symbolic structured form that will permit effective correlation, retrieval, and summary statement formation. (M)

(k) Recognize and associate similar content and thoughts, as well as similar symbols (identify relationships at phrase and sentence level). (H and M)

(l) Correlate new information with the existing corpus (internal to the system) in a dynamic adaptive manner. (M)

(m) Reduce redundancy from machine-stored information as corpus grows. (M)

(n) Select or reject content associations based upon criteria from a model of probable relevant correlations, depending upon the intended application of the system. (M)

(o) Logically derive explicit factual relations from input statements. (M)

(p) Derive implicit relations from input statements by analytical correlation methods:

(1) Discover existence of correlations (formal correlation). (M)

(2) Discover nature of correlations (causal correlation). (M)

(q) Associate and retrieve all correlated data in response to a stated request for specific information. (H and M)

(r) Convert correlated data to summary statements in syntactically, sensibly, logically, and factually correct sentences of natural language in response to a request for specific information (statement formation). (M)

(s) Request additional information, verification, or clarification from human analyst to assist in performing correlations. (H and M)

(t) Exercise executive control over its actions to perform related and sequential tasks efficiently and to respond to interrogations. (U) (M)

J. <u>Evaluation Tests</u>

Evaluation tests for the Fact Correlation System will be performed after the system has been essentially debugged; i.e., after the equipment has satisfactorily performed acceptance and interconnection tests and the programs written for the computer are operational. (U)

The evaluation of the system will be based upon its performance with respect to handling English language input, correlating the information contained in the input with any information already contained in memory, and responding to questions by using both the explicit and implicit information in its corpus. In terms of the flow chart in Figure 3-4, the test would evaluate the programs on the left together with the error routine as applied to these programs. The performance of the remainder of the system would illustrate only that the system was operational, a fact that will

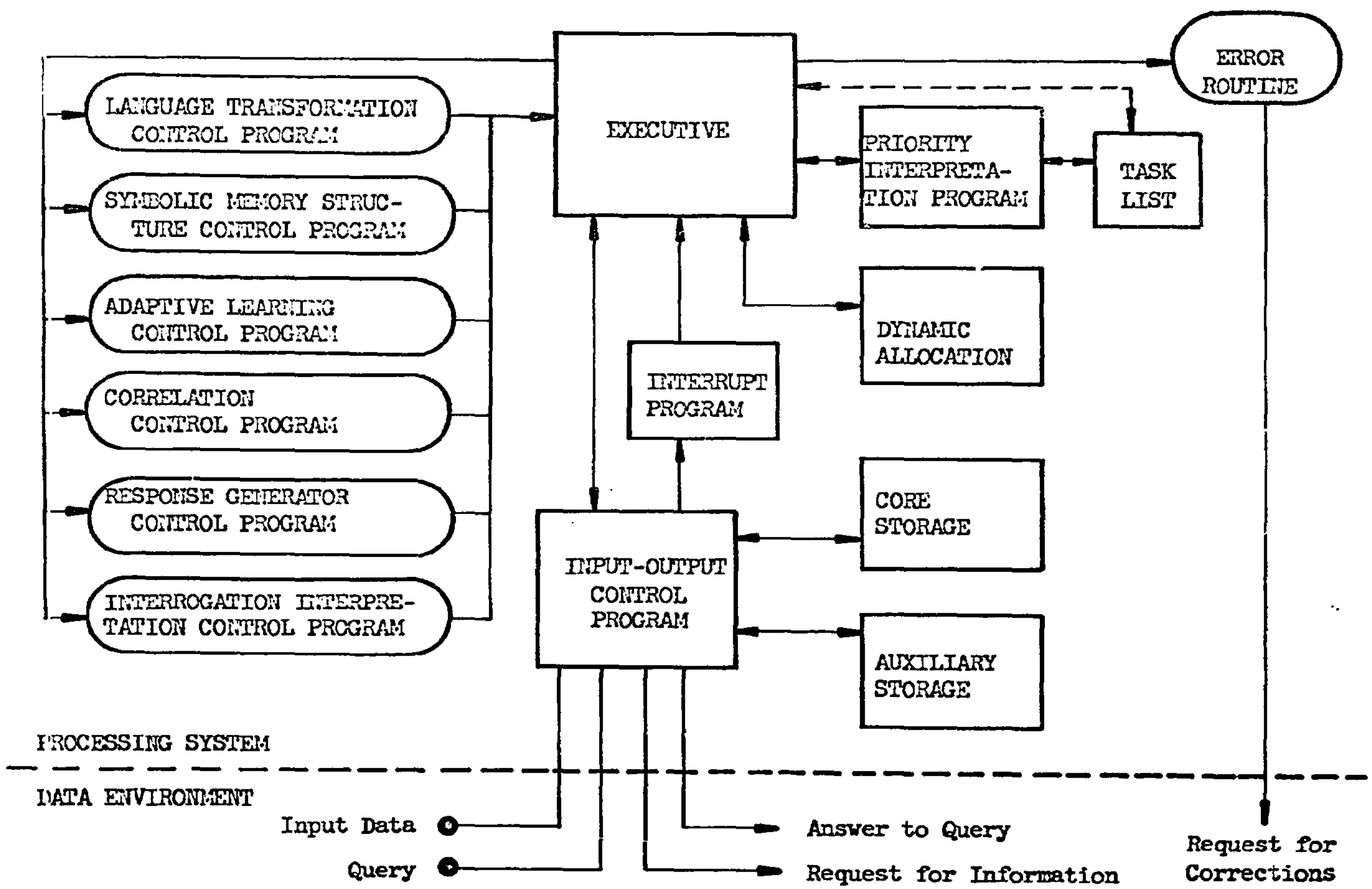

FIGURE 3-4. Programming System for Fact Correlation

have been established prior to the evaluation of system performance. (U)

The performance of the system will be tested by the results:

(a) The analysis of English sentences to determine the semantic structure of English.

(b) The synthesis of meaningful sentences in English.

(c) The logical structuring of information.

(d) The drawing of inferences about implicit information from a given set of explicit information.

The analysis of the English produced by the system can be compared with the analysis produced by human beings in applying the semantic theory to ordinary English. The criterion of validity will be that the sentences are acceptable, both syntactically and semantically, to a representative sample of users of the language. (U)

Assume that the computer has received information prior to this test. A dump of the memory will show what information the computer has already received and the structure it has attained. The following sample input (two paragraphs, describing the state of Wisconsin, taken from the Golden Book Encyclopedia, Number 16) is representative of information that may be used to evaluate system performance. Once the computer has analyzed and

WISCONSIN

This great lakes state is neither a large nor a crowded one. Wisconsin is the "Dairyland of America," but farming is not the only important work there. Fewer people live on farms than in cities. Wisconsin is a leading manufacturing state. It is also a much-enjoyed vacation land. Glaciers tore down the mountains of 25 million years ago. They left behind them a land of ridges, low rolling hills, valleys, and beautiful lakes.

Dairying is a good branch of farming in Wisconsin,

> where summers are not long and hot and much of the land is rolling or hilly. Most Wisconsin dairy farmers use many mechanical helpers to get big milk production. Milk is sent to city people or to dairy plants to be made into cheese, butter, and condensed or powdered milk.

stored the information according to the rules contained in the various programs for language transformation, memory structuring, adaptive learning, and correlation, an interrogator may ask a number of questions pertaining to these paragraphs. The computer will generate responses to the questions and may request additional information from the interrogator if necessary. (Of course, any question outside the frame of reference of the information received by the system will elicit either no response or a request for additional information.) (U)

A set of questions and anticipated replies based upon the sample paragraphs could be framed as follows:

<u>Question 1</u>: - What state is known as the "Dairyland of America"?

<u>Response 1</u>: - Wisconsin is the "Dairyland of America."

<u>Question 2</u>: - Why is dairying a chief industry of Wisconsin?

<u>Response 2</u>: - What means 'chief industry'?

<u>Statement 3</u>: - Chief industry same as important work.

<u>Response 3</u>: - Farming is important work. Dairying is a good branch of farming in Wisconsin, because summers are not long and hot and much of the land is rolling or hilly.

<u>Question 4</u>: - What are some dairy products?

<u>Response 4</u>: - Cheese, butter, and condensed or powdered milk are dairy products.

In Question 1 the interrogator asks a straightforward question. The answer illustrates the computer's ability to store and retrieve information properly. Response 1 is a proper reply. (U)

In Question 2 the interrogator uses the term 'chief industry.' Response 2 indicates that the computer has no knowledge of the meaning of 'chief industry.' Therefore, the interrogator must define 'chief industry' in terms the computer understands. In Statement 3, the human response to Response 2, 'chief industry' is equated to 'important work.' The computer is then able to reply in Response 3 after several correlations of data. First it finds that "...farming is not the only important work...," which it analyzes to obtain "farming is important work" since the only restricts the negation sense of not. This fact, however, does not answer the question that pertains to dairying. Further correlation and analysis show "Dairying is a good branch of farming...," which indicates a relation between dairying and farming. The remainder of the sentence about dairying gives a reason for that type of farming in Wisconsin; namely, "...summers are not long and hot and much of the land is rolling or hilly." This statement answers the question, "Why?" and the computer is now able to respond to the original question. (U)

In order to answer Question 4, the system must correlate dairy products with things made from milk. The next to last sentence in Paragraph 2 states that "...dairy farmers...get big milk production." Product is the root word of production, hence milk products are the products of dairy farmers. The last sentence lists some of the things produced from milk--hence, some milk products. These things are selected to answer the question "What are some dairy products?" and the answer is shown in Response 4. (U)

This type of input and questioning would illustrate whether or not the system is successfully transforming the input, structuring the memory, correlating the proper items, and generating meaningful sentences. The adaptive learning process is carried on as the computer receives more information to analyze. A question may provide useful information. For example, to test whether the computer has learned the meaning of chief industry, the following statement and question could be presented to the system:

Statement: - In Florida raising citrus fruit is important work.

Question: - What is a chief industry of Florida?

The computer should be able to respond with the following statement:

Response: - Raising citrus fruit is a chief industry of Florida. (U)

The error routine may be tested to prove its operational capability in detecting errors in input data and reporting them to an analyst or interrogator. For example, if the first sample question had read "What states is the Dairyland of America?" an error should have been indicated. The question seems to be requesting a plural answer for states, although the singular verb is is used with a singular descriptive name Dairyland of America. The error routine should ask for a verification as follows:

Response: - Is more than one state the Dairyland of America?

This test also demonstrates the ability of the program to recognize violations of syntax. (U)

Another example could be used to illustrate the ability of the program to recognize logical consistency. If the statement, "Wisconsin is one of the largest and most populated of the 50 states," were given to the computer as new information, the system should recognize a contradiction and indicate it as follows:

<u>Conflict</u>: - Wisconsin is neither large nor crowded.

- Wisconsin is one of the largest and most populated of the 50 states.

- Which statement is correct?

However, if the computer were given the information that "Trenton is the capitol of Wisconsin," it would not be able to recognize whether the statement is true or false. The system would accept the statement as true unless it contradicted some previously received information. (U)

Many tests will be needed in order to obtain a general measure of system effectiveness. The design of these tests is an important function of system analysis. By studying the way in which the computer correlates certain information, it may be possible to determine methods of improving these techniques. If improper correlations are made or if the language transformation program is inadequate, these programs will have to be revised. A well conceived test program may not only evaluate the performance of the system, but also provide clues for improving processing operations. (U)

K. Summary

Part III presents a concept for developing a Fact Correlation System. The essence of this concept is that an adequate Fact Correlation

System can be developed only by maintaining the system viewpoint consistently. This method entails adequate definitions of user requirements and system objectives. From these definitions system tasks are defined and the means of implementing them discovered and developed. The crucial aspects of this procedure are twofold:

(a) Consider any task within the appropriate frame of reference as it relates to other tasks, the entire system, and the system objectives.

(b) Consider at all times the three basic elements of the system; equipment, computer programs, and personnel. This consideration emphasizes the parallel development of the capabilities of each element with constant attention to improved performance that may be obtained by taking full advantage of the complementing effect of the interaction among these elements and the unique capabilities of each. (U)

Most attempts to solve the kind of problems encountered in developing a Fact Correlation System contrast with this concept. Existing research tends to take some of the elements of such a system as given or else to ignore them entirely. This methodology has resulted in isolated, seemingly unrelated results and piecemeal solutions that fail to solve satisfactorily the complex problems involved. This situation can be remedied by conducting research and development within the scope of a unified, integrated system approach. (U)

IV. RESEARCH PLAN

A. Research Objectives

The research objectives for a Fact Correlation System have been specified implicitly in other parts of this report. They are given specifically in terms of system capabilities at the end of Part III. Operational capabilities will be presented in Section C. There is no research involved directly in the system analysis, the establishment of system utilization procedures or system test and evaluation. Some research in computer programming may be required. If so, it will probably be in the area of methods of storing, comparing and relating data, in efficient indexing and manipulating of indices, and in integrating and selecting combinations of programs to perform various tasks. The major areas of research are equipment requirements and linguistic-learning correlation techniques that will be converted into computer programs. (U)

The objective in equipment research is to develop performance specifications for equipment that will be effective in processing linguistic data and will permit the optimal operation of the necessary computer programs, including responses to human queries, when operating as a system. Consequently, interaction between equipment components is stressed as an area of analysis. (U)

The objectives of research in techniques are: to develop practical machine learning techniques, data structuring techniques, and linguistic analysis and synthesis techniques; and to apply these techniques to the correlation and retrieval of internally stored factual information in order to respond to a wide variety of human interrogations entered into the system in natural language. These techniques must be developed to the point where they can be converted into computer programs for the processor of the system. (U)

B. <u>General Research Concepts</u>

There are three basic approaches to the design of systems. These approaches are:

(a) The <u>sequential</u> subsystem design approach.

(b) The initial-ultimate (<u>evolutionary</u>) operational capability approach.

(c) The combination of these two approaches. (U)

The first approach is usually adopted for the design of systems whose requirements and modifications are well known and where the design methodology for equipment and techniques is well established. The approach consists of the design and construction of subsystems or devices that perform subfunctions on a relatively independent or sequential basis. This method is a traditional engineering design procedure, and it is most effective for a system that represents a relatively modest advancement in the state-of-the-art of a well-established technology. For example, this approach is used in designing and manufacturing automobiles, electronic household appliances, and typewriters. This method can be used only when the operation of the subsystems is relatively independent, or when their interrelationships are well defined <u>a priori</u> and modifications to the system design will be neither frequent nor extensive. (U)

The second approach is most often used for systems that involve entirely new operational concepts and where extensive research may be required to establish functional interrelationships within the system. This approach consists of establishing an initial operational capability and an ultimate operational capability. The design of the entire system is then governed by the objectives of the relatively rudimentary requirements for initial operational capability. Research into new techniques to provide more sophisticated system capability takes place concurrently with the development of the initial system. When initial operational capability has been achieved, the entire system is modified and upgraded to a more sophisticated capability level in accordance with the latest developments in techniques and equipment. This process is then repeated until a performance level consistent with the ultimate system objectives is achieved. This evolutionary method has been used in the development of some complex weapons systems and command-control systems. (U)

A third approach to system design and development consists of the combination of the sequential and evolutionary methods. This approach is ideal for attaining maximum system capability in the shortest time. This goal can only be accomplished, however, if system development is effectively managed and implemented. The management of this type of research and development effort is more difficult and complex than either of the first two methods. (U)

An effective automated Fact Correlation System involves entirely new data processing concepts and man-machine interactions and will represent a major advance in the state-of-the-art of information storage, correlation, and retrieval. Since the development of such a system requires considerable research (as discussed elsewhere in this report), and since system functions and their interactions are not completely known at this time, the first approach to system design cannot be effectively applied. Even if it could be, other disadvantages preclude its use. For example, one of the requirements for the Fact Correlation System is to develop or implement intermediate operational products that perform some system functions independently of other system functions. The first design approach does result in the development of such intermediate subsystems. However, it requires approximately 10 to 25 percent additional effort to convert these subsystems into independent operational entity. In addition, insofar as the subsystems are not independent, when the design of two or more of them has been completed, considerable effort may be required to modify their design so that they can function as an integral part of a larger subsystem. Another difficulty in applying this method to a research contract is that short-term progress is usually quite difficult to measure and monitor. (S)

In contrast to the sequential approach, the evolutionary design approach produces intermediate products that are complete operational systems. These intermediate systems will have considerably less capability than the system produced by the sequential approach. However, the capability of the ultimate system developed by the evolutionary approach will at least equal and probably exceed that of the system developed by the first method. The evolutionary approach also permits gradual investment, appropriately phased training of personnel, and the introduction of new procedures. This method should be adopted as the primary approach to develop an automated system for correlating factual information. (U)

In the development of a Fact Correlation System the third method may be used to the extent that it is practical and feasible. These conditions are determined by the degree to which the system functions and tasks can be detailed and their dependence or independence established. Then appropriate analytical techniques may be employed in phasing and scheduling the system development effort. The research and development effort will be described by complex relationships among series and parallel tasks that complicate attempts to simplify planning procedures. (U)

In performing research and development for a Fact Correlation System, there is no assurance of completing a given task by a certain time; There is only a likelihood. To ignore this uncertainty is to reduce the realism of any program planning. Since uncertainty is a characteristic of this type of research, it is desirable to forecast areas of uncertainty in advance, if effective anticipatory action is to be planned. An important technique that aids in anticipating difficulties and allocating resources for optimal effort is the so-called "critical path" method of project scheduling, which is discussed in detail in the next section. (U)

C. Research Schedule

The development of a useful research schedule requires a reasonably comprehensive knowledge of system functions and tasks and their interrelationships. The major elements of the development job for the Fact Correlation System are shown in Figure 4-1. Five major areas of effort are shown in the left margin. Each of these areas is subdivided into the tasks shown in the figure. The abscissa represents calendar time from project start to system delivery. The length of the blocks indicates in

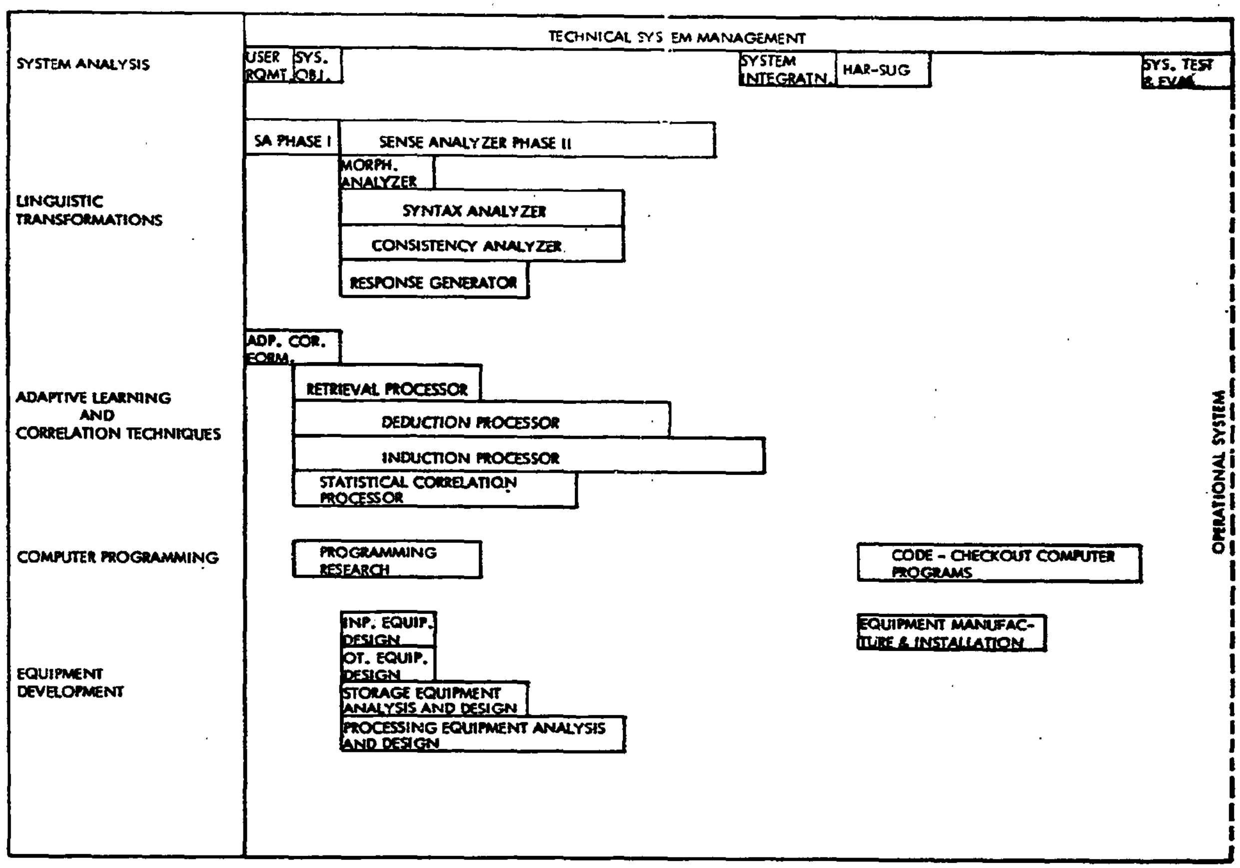

FIGURE 4-1. FACT CORRELATION SYSTEM DEVELOPMENT TASKS

general the approximate phasing and relative time required to complete each task. The location of the blocks indicates the sequential relationship among tasks in each area. The tasks in this figure are the major elements of the development job for the Fact Correlation System. These tasks must be accomplished regardless of the system design method that is adopted. Since this figure is only intended to be schematic, the relationships among tasks within the five different areas are not shown. If there were no such relationships, this figure would be a general outline of the sequential approach. A more precise relationship between system tasks is illustrated in Figure 4-3. (U)

Figure 4-2 is an outline of the second method of approach. It shows five levels of system capability beginning with an initial operational capability A and ending with the ultimate capability E. It should be emphasized that each of the letters A through E represents an operational Fact Correlation System. An increase in the capability of the elements of each system is indicated by a change in the Roman numerals, which may be interpreted as model numbers. The increasing length of the vertical bars shows the relative improvement in system capability from one model to the next. The general capability for the initial and ultimate systems can be specified now. The capability of intermediate system models, however, is extremely difficult to specify because such capability depends upon the research problems encountered in each task and the success and complexity of the particular techniques developed to solve them. The projected goals for the initial and ultimate operational capability of an automated Fact Correlation System are given below:

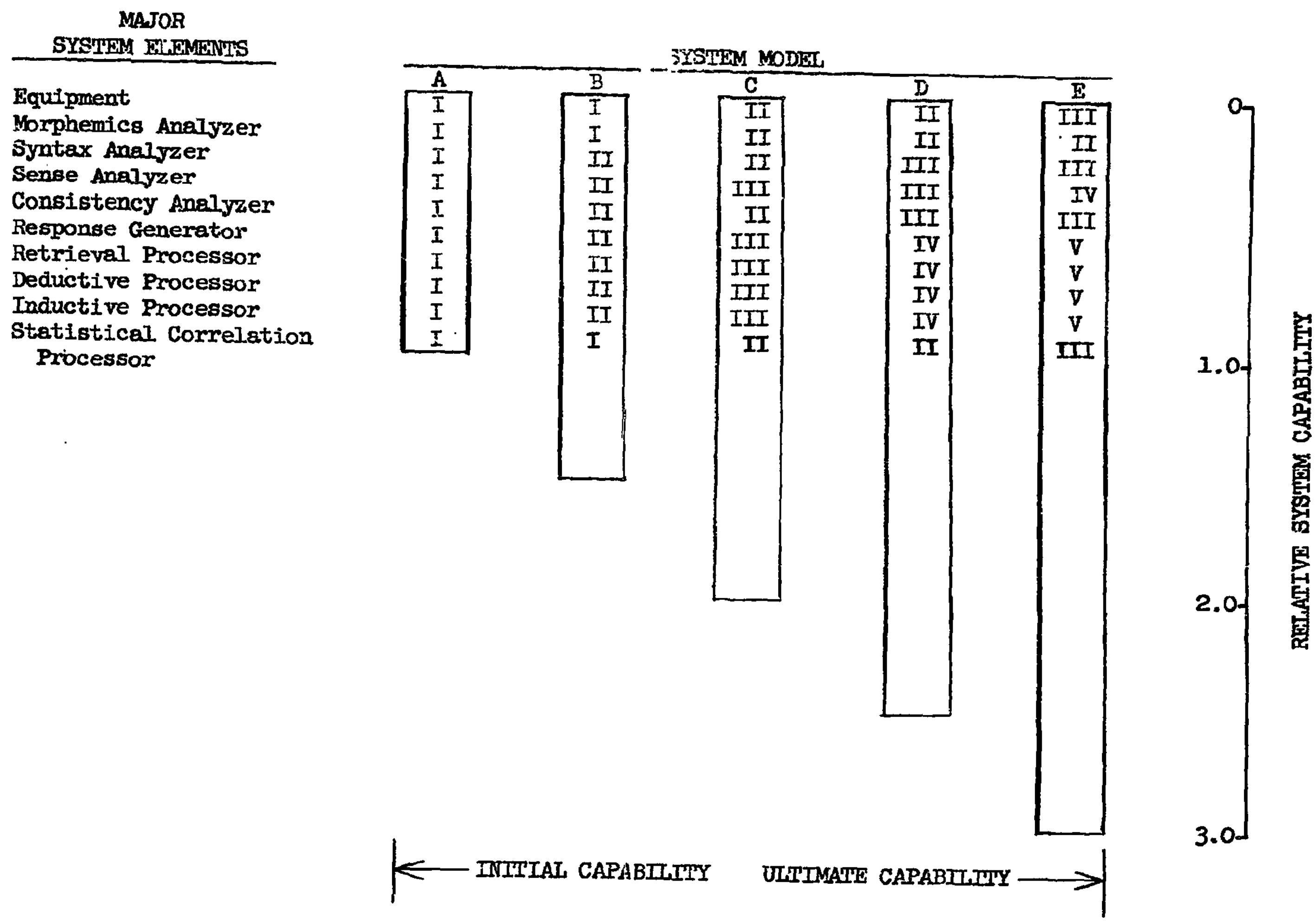

FIGURE 4-2. Development of System Operational Capability

1. <u>Initial Operational Capability</u>

(a) Inputs will be restricted to simple English sentences; that is, sentences consisting of a single clause.

(b) The syntax analyzer will be capable of phrase and sentence recognition and analysis.

(c) The sense analyzer will check the validity of input sentences and construct in computer memory the sense structure of the simple sentences of the language from the inputs received.

(d) A dictionary consisting of key, operational, and structure words of the language will be constructed in computer memory.

(e) A rudimentary data structuring system, which permits convenient correlation and retrieval of information, will be developed within the computer.

(f) The computer will accept human interrogations in the form of formatted simple questions; that is, questions consisting of a single clause.

(g) The computer will perform primitive adaptive learning operations to correlate structured data.

(h) The computer will generate and print simple sentence responses to human interrogations. (U)

2. <u>Ultimate Operational Capability</u>

(a) The system will accept unrestricted English sentences as input.

(b) The system will analyze syntax of input sentences to whatever degree necessary to perform any other system functions.

(c) The system will analyze all input sentences for sense validity and construct the sense structure of the language from the input sentences received.

(d) The system will check all input sentences for logical consistency.

(e) The system will interrogate human analysts with questions in natural language.

(f) The system will accept unformatted human interrogations.

(g) The system will organize and memorize the factual content of input statements. The statements will be stored in a symbolic structured form that will permit sophisticated correlation and retrieval of summary statement information.

(h) The system will correlate new information with the existing corpus in a dynamic adaptive manner.

(i) The system will reduce redundancy from machine-stored information on the basis of relative frequency of use with time.

(j) The system will logically derive explicit factual relations from input statements.

(k) The system will derive implicit factual relations from input statements by analytical and statistical methods.

(l) The system will associate and retrieve all correlated data in response to a human request for specific information.

(m) The system will respond to human interrogation by printing natural language sentences containing summary information.

(n) The system will request additional information, verification, or clarification from the human analyst to assist it in performing correlations. (U)

The method of scheduling these project activities will now be discussed in some detail. It is important to state at the outset that the following discussion and analysis is primarily illustrative. The use of the following techniques for realistic project scheduling requires a much more detailed breakdown of relationship among system development tasks, a detailed study of manpower requirements and allocation, and the introduction of cost factors as a function of development time. It will also require frequent updating of time estimates and the scheduling network. Nevertheless, a few important conclusions can be drawn from the relatively crude analysis that follows. (U)

Although several specific scheduling programs of the type to be discussed have been developed, the general descriptive name of critical path scheduling will be used in this report. The foundation of critical path scheduling is to establish the flow chart or network of project milestones and check points that must be attained if the over-all program is to be completed. A network for a Fact Correlation System with ultimate operational capability is shown in Figure 4-3. Project activities or tasks are indicated by arrows. For a given activity (represented by an arrow emanating from a circle) to begin, the tasks represented by all the arrows terminating at that circle must have been completed. When the milestone network has been completed, time estimates for accomplishing tasks are superimposed upon it. (U)

In order to use statistical techniques concerning the likelihood of completion by a given time, three time estimates are made for each task. These subjectively determined estimates are: an optimistic time, a pessimistic time, and a most likely time. Although the estimates are subjective, they should be made by personnel highly experienced in each task area. These times are then combined into a single weighted average time for each activity. Some estimates for the network in Figure 4-3 are shown in Table 4-1. These times may be interpreted as absolute or relative times with certain limitations. The critical path is then defined as the longest path through the network in terms of the sum of the weighted average times for each distinct path. Each activity on the critical path is a critical activity. An activity is critical if a slippage in its completion time will result in a slippage in the system delivery date. Slack time for a

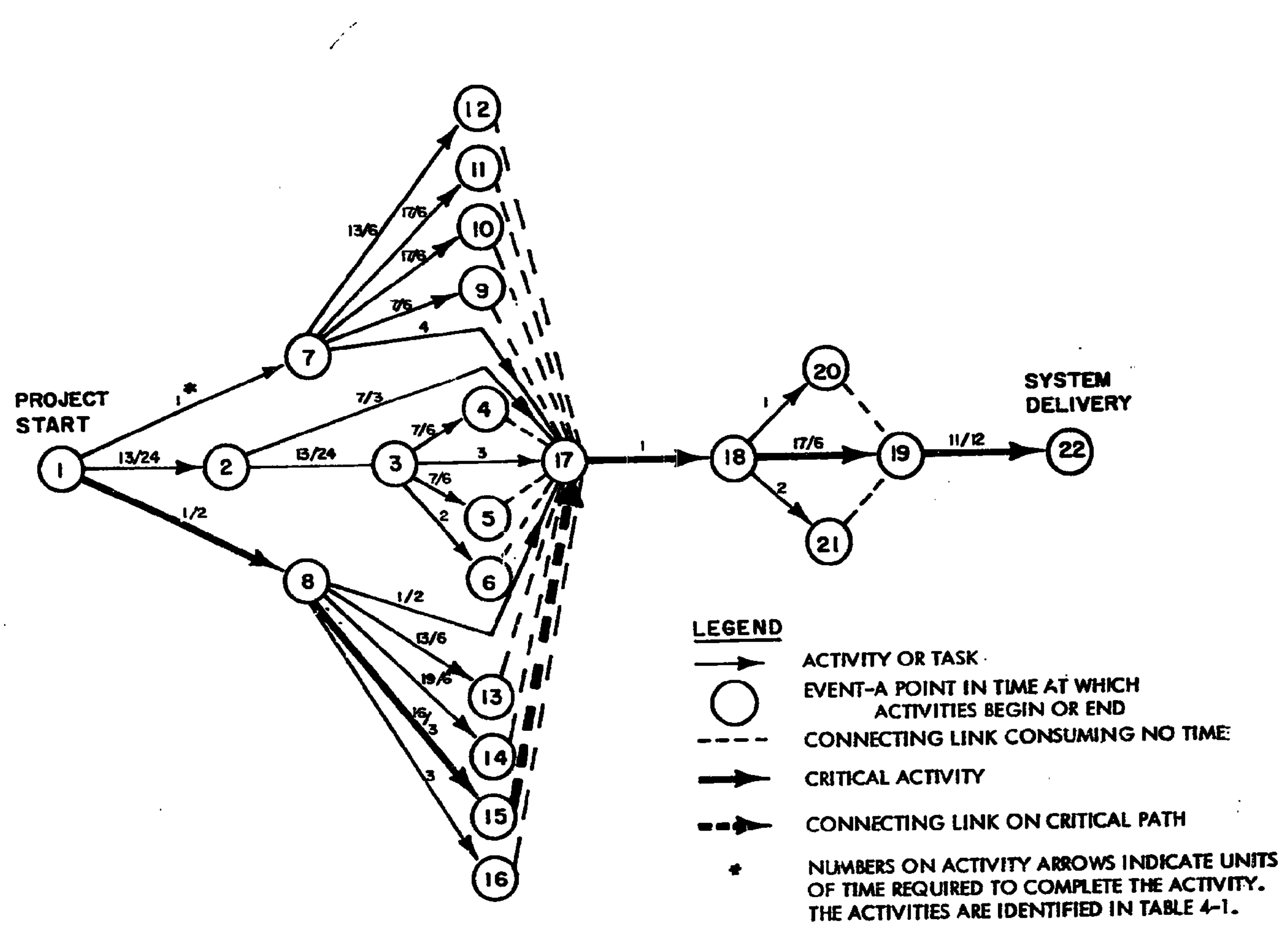

FIGURE 4-3. TYPICAL MILESTONE NETWORK FOR A FACT CORRELATION SYSTEM

TABLE 4-1. (PART I). PROGRAM ACTIVITIES, DESIGNATIONS, AND COMPLETION TIME ESTIMATES

ACTIVITY	DESIGNATION IN FIGURE 4-3	OPTIMISTIC COMPLETION TIME	MOST LIKELY COMPLETION TIME	PESSIMISTIC COMPLETION TIME	WEIGHTED* AVERAGE COMPLETION TIME
User Requirements	1-2	1/4	1/2	1	13/24
System Objectives	2-3	1/4	1/2	1	13/24
Morphemics Analyzer	7-9	1	1	2	7/6
Syntax Analyzer	7-10	1	3	4	17/6
Sense Analyzer Phase I	1-7	1/2	1	3/2	1
Sense Analyzer Phase II	7-17	3	4	5	4
Consistency Analyzer	7-11	1	3	4	17/6
Response Generator	7-12	2	2	3	13/6
Adaptive Correlation Requirements Phase I	1-8	1/4	1/2	1	1/2
Adaptive Correlation Requirements Phase II	8-17	1/4	1/2	1	1/2
Retrieval Processor	8-13	2	2	3	13/6
Deduction Processor	8-14	2	3	5	19/6
Induction Processor	8-15	4	5	8	16/3
Statistical Correlation Processor	8-16	2	3	4	3

$$* \ WA = \frac{O + 4ML + P}{6}$$

TABLE 4-1. (PART II). PROGRAM ACTIVITIES, DESIGNATIONS, AND COMPLETION TIME ESTIMATES

ACTIVITY	DESIGNATION IN FIGURE 4-3	OPTIMISTIC COMPLETION TIME	MOST LIKELY COMPLETION TIME	PESSIMISTIC COMPLETION TIME	WEIGHTED* AVERAGE COMPLETION TIME
Advanced Programming Techniques	2-17	2	2	4	7/3
Input Equipment Analysis and Design	3-4	1	1	2	7/6
Output Equipment Analysis and Design	3-5	1	1	2	7/6
Storage Equipment Analysis and Design	3-6	1	2	3	2
Processing Equipment Analysis and Design	3-17	2	3	4	3
System Integration	17-18	1/2	1	3/2	1
HAR's and Utilization Guides	18-20	1/2	1	3/2	1
Coding and Checkout of Computer Programs	18-19	1	3	4	17/6
Manufacture and Installation of Equipment	18-21	1	2	3	2
System Test and Evaluation	19-22	1/2	1	1	11/12

* $WA = \frac{O + 4ML + P}{6}$

given task is defined as the difference between the latest time it may be completed without affecting the system delivery date, and the time it is expected to be completed. Expected time of completion for a given task is the sum of the weighted average times of all tasks that must be completed before the given task can begin plus the weighted average time required to complete the given task itself. For non-critical activities the slack time can then be determined. (U)

Critical activities and non-critical activities with their slack times are shown in Table 4-2. By definition, critical activities have zero slack time. Examination of the critical activities shows that in the early part of the system development effort the formulation and development of adaptive and inductive machine techniques are critical. In the middle period system integration is critical. In the latter stages computer programming implementation and checkout, and system test and evaluation are critical. These conditions accord with experience in the development of other man-machine systems. Examination of the non-critical activities verifies the intuitive judgment that the development of equipment is not critical for a Fact Correlation System. The early phases of system analysis and the development of system utilization procedures are not critical either. The fact that the language transformation tasks are not critical probably reflects the current situation; namely, that more is known (and more research performed) about the machine analysis of language than about machine learning and inference. (U)

Implicit in the completion time estimates used as schedule inputs is a consideration of the use of resources and of the performance

TABLE 4-2. CRITICAL ACTIVITIES AND NON-CRITICAL ACTIVITIES WITH SLACK TIMES

ITEM	CRITICAL PATH ACTIVITIES	SLACK TIME
1.	Formulate Adaptive Correlation Requirements (Phase I)	0
2.	Develop Induction Processor	0
3.	Integrate System	0
4.	Code - Checkout Computer Programs	0
5.	Test and Evaluate System	0
	NON-CRITICAL ACTIVITIES	
1.	Develop User Requirements	7/4
2.	Establish System Objectives	7/4
3.	Develop HAR's and Utilization Guides	11/6
4.	Develop Morphemics Analyzer	11/3
5.	Develop Syntax Analyzer	2
6.	Develop Sense Analyzer (Phase I)	5/6
7.	Develop Sense Analyzer (Phase II)	5/6
8.	Develop Consistency Analyzer	2
9.	Develop Response Generator	8/3
10.	Formulate Adaptive Correlation Requirements (Phase II)	29/6
11.	Develop Retrieval Processor	19/6
12.	Develop Deduction Processor	13/6
13.	Develop Statistical Correlation Processor	7/3
14.	Perform Input Equipment Analysis and Design	43/12
15.	Perform Output Equipment Analysis and Design	43/12
16.	Perform Storage Equipment Analysis and Design	11/4
17.	Perform Processing Equipment Analysis and Design	7/4
18.	Manufacture and Install Equipment	5/6
19.	Develop Advanced Programming Techniques	71/24

characteristics of the system. The question then arises: can resources be re-allocated or performance modified to shorten the critical path appreciably? (U)

One of the purposes in adopting the design approach of providing a sequence of systems for fact correlation is to upgrade system performance by modest increases in relatively shorter periods of time. This technique will help to reduce the project development time by permitting more realistic assessment and establishment of performance objectives and capability. (U)

Since equipment development is not critical in this project, the primary resources are personnel and computer time. Use of computer time will be optimized and probably will not be critical. Re-allocation of personnel in a research project is a complex problem involving personnel skills, costs versus time saved, and the probability of saving a certain amount of time by adding a certain number of personnel to a task. Critical tasks 1 and 2 in Table 4-2 are complex research tasks. For this reason the variance of completion time estimates is large; moreover the type of product that will result cannot be anticipated with much reliability. The allocation of more man-power to this type of task usually will not lower completion time significantly. However, the allocation of more man-power for the fourth critical task, computer programming, may help to reduce completion time significantly. Completion times for critical tasks 3 and 5 are relatively brief. Although additional man-power applied to these tasks might reduce completion times by 10 or 20 percent, the reduction in the length of the critical paths is smaller than the accuracy of the time estimates. Consequently, such re-allocations do not seem worth considering. (U)

These remarks are summarized in Table 4-3. Unless the answer to both of the questions in Columns 2 and 3 is, "yes," it is not worthwhile re-allocating man-power in an attempt to reduce total system development time. At this time, it appears that computer programming is the only area where added man-power may reduce development time significantly. It may also be feasible to begin the programming effort before the system integration is completed. This mode of scheduling would also reduce total system development time considerably. Another consideration in estimating computer programming time is the match between equipment capabilities and programmed capability. When functions to be automated have been reduced to algorithms, the specific computer operations that are necessary can be built into the equipment or programmed. If more special purpose functions are built into the equipment, the equipment development time is longer, but the programming time is shorter. This match or trade-off must be decided at the appropriate time in system development. Then, more reliable time estimates can be made. (U)

It is appropriate to reiterate that a more detailed task breakdown and scheduling network may well indicate the feasibility of beginning the computer programming earlier. It may also reveal other areas where development time may be reduced, and it could even result in a somewhat different critical path. The problem remains, however, that completion times and performance capabilities resulting from complex research task efforts are difficult to specify. This problem is best handled by the phased development of intermediate systems (and capabilities) and the continual updating of the scheduling network and completion time estimates. Short term

TABLE 4-3. RE-ALLOCATION OF MAN-POWER TO CRITICAL ACTIVITIES

CRITICAL ACTIVITY	WILL ADDITIONAL MAN-POWER AFFECT COMPLETION TIME SIGNIFICANTLY?	WILL SHORTER TASK COMPLETION TIME REDUCE CRITICAL PATH SIGNIFICANTLY?	RE-ALLOCATION EFFECTIVE?
1. Formulate Adaptive Correlation Requirements (Phase I)	No	No	No
2. Develop Induction Processor	No	Yes	No
3. Integrate System	Yes	No	No
4. Design - Code - Checkout Computer Programs	Yes	Yes	Yes
5. Test and Evaluate System	Yes	No	No

re-allocation of man-power will be more effective on this basis. (U)

Estimated man-power requirements for the Fact Correlation System are provided in Table 4-4. These estimates are based on the development of the Fact Correlation System to an ultimate operational capability. The task breakdown is essentially the same as in Figure 4-1. The total number of years for each task was taken from the most likely completion time estimate of Table 4-1. In addition, Table 4-4 shows the number of men and total man-years for each task. This information is also totaled for each of the five major task areas and for the entire development effort with the exception of equipment manufacture and installation. Man-power estimates for technical management have also been included. These estimates are based upon one technical supervisor for each of the five task areas for the duration of the task. (U)

D. Personnel Requirements

The types of personnel required for each task area will be discussed briefly. The personnel requirements for the System Analysis task area are more closely related to a systems point of view and a talent for system analysis rather than any particular academic background. For the sake of balance, this area will probably include personnel with backgrounds in the natural sciences, mathematics, and the social sciences. Considerable experience in the analysis of data processing systems is a prerequisite for personnel in this task area. The Linguistic Transformations task area will include personnel with backgrounds in linguistics, mathematics, and logic. The personnel in the Adaptive Learning and Correlation Techniques area will be psychologists, mathematicians, and logicians. Personnel for

TABLE 4-4. (PART I). MAN-POWER REQUIREMENTS FOR FACT CORRELATION PROJECT

TASKS	NUMBER OF MEN	TOTAL NUMBER OF YEARS	TOTAL MAN-YEARS
A. SYSTEM MANAGEMENT AND ANALYSIS	17-22	---	27-32
1. Technical Management	5	---	17
2. User Requirements	2	1/2	1
3. System Objectives	2	1/2	1
4. System Integration	3-5	1	3-5
5. HAR's and Utilization Guides	1-2	1	1-2
6. System Test and Evaluation	4-6	1	4-6
B. LINGUISTIC TRANSFORMATIONS	6-10	---	19-33
1. Morphemics Analyzer	1	1	1
2. Syntax Analyzer	1-3	3	3-9
3. Sense Analyzer	2-3	5	10-15
4. Consistency Analyzer	1-2	3	3-6
5. Response Generator	1	2	2
C. ADAPTIVE LEARNING AND CORRELATION TECHNIQUES	12-15	---	32-46
1. Formulation	4	1	4
2. Retrieval Processor	2	2	4

TABLE 4-4. (PART II). MAN-POWER REQUIREMENTS FOR FACT CORRELATION PROJECT

TASKS	NUMBER OF MEN	TOTAL NUMBER OF YEARS	TOTAL MAN-YEARS
3. Deduction Processor	2-3	4	8-12
4. Induction Processor	2-4	5	10-20
5. Statistical Correlation Processor	2	3	6
D. COMPUTER PROGRAMMING	14-17	---	40-49
1. Programming Research	2	2	4
2. Code - Checkout Computer Programs	12-15	3	36-45
E. EQUIPMENT DEVELOPMENT	7-8	---	13-16
1. Input Equipment Analysis and Design	2	1	2
2. Output Equipment Analysis and Design	1	1	1
3. Storage Equipment Analysis and Design	2	2	4
4. Processing Equipment Analysis and Design	2-3	3	6-9
5. Equipment Manufacture and Installation	*	2	*
TOTAL	56-72	---	131-176

* Estimates for this task are best obtained from manufacturers of specific types of equipment and from the cost of existing equipment.

the Computer Programming area will be of two types. Research programmers for task 1 and programmers for operational and executive programming under task 2. Personnel for research in advanced programming techniques must have experience and facility in the general theory of programming and programming languages. Requirements for operational programmers are less stringent. Experience and "know-how" in both areas are more important than a particular academic background except in so far as it is computer oriented. Personnel for equipment analysis, design, and development will be electronic engineers. (U)

V. CONCLUSIONS

This report presents the results of a study of current techniques and equipment and an evaluation of the required areas of research in techniques, computer programming, and equipment necessary for the development of an operational Fact Correlation System capable of correlating intelligence information automatically. This volume of the report surveys the system requirements in general and develops a methodology and plan for an applied research program. The second volume contains a detailed review of existing techniques and their applicability to the system requirements. (U)

In a study of this type it is inevitable that the definitiveness and verifiability of the conclusions vary widely. Some are quite tentative; others are firm, verified by professional knowledge and the results of related studies. (U)

The conclusions derived from this study are:

1. An extensive continuing system analysis is necessary for the development of an automated Fact Correlation System. An adequate definition of user requirements and system objectives is essential. (See Volume 1, Part III.) (U)

2. The system must be capable of accepting and analyzing natural language input sentences automatically. These processes must recognize well-formed constructions and detect and resolve ambiguities at the levels of morphology, syntax, semantics, and logical and factual consistency. The system must also be capable of generating sensible, consistent

sentences in natural language in response to queries for specific information. (See Volume 2, Part III.) (U)

3. Current linguistic transformation techniques in the areas of morphology, syntax, and logical consistency are generally adequate or nearly adequate. (See Volume 2, Part II.) (U)

4. Current techniques for linguistic analysis and synthesis and for ambiguity resolution in the areas of semantics and factual consistency are in a primitive state. Extensive research will be required in these areas. (See Volume 2, Part III.) (S)

5. The system must be able to organize and structure input information after automated linguistic transformation in order to derive explicit and implicit factual relationships. (See Volume 2, Part III.) (U)

6. The system must contain adaptive elements that enable it to associate similar content, classify information, and select only the relevant structural, operational, and functional relationships. The system must be capable of improving its performance with time. (See Volume 2, Part III.) (U)

7. The system must have the capacity to perform logical deductions and inductions and, possibly, statistical correlations. (See Volume 2, Part III.) (U)

8. The system must provide for extensive man-machine interaction. The system must permit frequent queries from man to machine and form machine to man. (See Volume 2, Part III.) (U)

9. Current research in neuron networks, probabilistic self-organizing systems, and machines like the perceptron is not applicable to a Fact Correlation System because:

(a) Such systems begin with no information whatever.

(b) These systems do not have enough rules for self-organization.

(c) These techniques are not concerned with relating linguistic data but with simulating elementary and fundamental learning processes.

(See Volume 2, Part III.) (U)

10.. Research in game playing and theorem proving or problem solving machines is not useful in fact correlation except for the heuristic concepts involved. Such systems have too many rules. For specific game situations all the rules are known; in many cases, the strategies and tactics are also known a priori. Consequently, too much human knowledge and experience is built into the system for a specific situation instead of a more general rational methodology for forming associations and relations among concepts. (See Volume 2, Part III.) (U)

11. The executive control program plays a central role in a Fact Correlation System because:

(a) The requirement for man-machine exchange of information during processing.

(b) The exact composition, organization, and sequencing of subprograms may be unknown prior to execution.

(c) Several tasks may be executed at the same time.

(See Volume 2, Part IV.) (U)

12. The effectiveness for fact correlation of various programming languages and systems such as list processors and other problem-oriented languages as well as multiple or parallel programming techniques should be investigated. (See Volume 2, Part IV.) (U)

13. An input system consisting of keypunched cards entered into the system through a card reader is an unacceptable bottleneck for a Fact Correlation System with a high reception rate of raw input data. (See Volume 2, Part V.) (U)

14. The following equipment configurations provide a reasonable balance between input entry time and processing time for a large and complex Fact Correlation System:

(a) For a high input rate, one reading machine and at least 64K core storage with disc files as auxiliary storage. Processing and storage access times are based upon the fastest commercial computers presently available.

(b) For lower input rates, one reading machine and 32K core storage with magnetic tapes as auxiliary storage.

(c) A second choice for lower input rates is an optical scanner with 32K core storage and disc files. Total input entry and processing time for this configuration is larger than for configuration (b).

(See Volume 2, Part V.) (U)

15. At present, text reading machines under development are hindered by their inability to handle paper at sufficiently high speeds. Improvements in the reliability of these devices and in their ability to read degraded print or handwritten characters can be expected in the near future. (See Volume 2, Part V.) (U)

16. A parallel random access memory that can retrieve information in one memory cycle regardless of memory size would be a major equipment breakthrough for a Fact Correlation System. (See Volume 2, Part V.) (U)

17. Microminiature elements, thin films, and the application of new physical principles will provide increased equipment capability for fact correlation at decreased costs. (See Volume 2, Part V.) (U)

18. Other developments in processing and storage equipment will improve system performance significantly until techniques are developed to a much greater degree than at present. This development of techniques is necessary in order to utilize equipment capabilities more fully. (U)

19. The design and development of the system is best accomplished by an evolutionary approach that results in a sequence of operational systems of increasing capability until a satisfactory ultimate operational capability is attained. (See Volume 1, Part IV.) (U)

20. The following tasks will be critical in determining the total system design and development time:

(a) Formulation of Adaptive Correlation Requirements (Phase I.)

(b) Development of Induction Processor.

(c) System Integration.

(d) Coding and Checkout of Computer Programs.

(e) System Test and Evaluation.

(See Volume 1, Part IV.) (U)

21. At present, a reduction in the time required for the completion of critical tasks by re-allocating man-power is foreseen only for the coding and checkout of computer programs. No appreciable time can be saved by increasing the number of personnel performing research in techniques. (See Volume 1, Part IV.) (C)

VI. RECOMMENDATIONS

Most of these recommendations are implicitly or even explicitly stated in the conclusions listed in Part V. The following recommendations provide somewhat more specific information concerning the suggested over-all approval and specific activities derived from the conclusions. (U)

1. System Planning and Analysis

Adopt an evolutionary approach to the design and development of a Fact Correlation System and maintain an extensive continuing system analysis activity. This method will result in a sequence of operational systems of increasing capability until an adequate operational capability is achieved. In conjunction with this approach, maintain a critical path schedule to ensure optimal allocation of resources and concentration on the critical tasks for each operational system. (U)

2. Linguistic Transformations

Develop techniques that will automatically analyze and synthesize input sentences, recognize well-formed constructions, and resolve ambiguities. These operational processes should include the levels of morphology, syntax, semantics, and logical and factual consistency. Use existing techniques or modifications of them to the maximum possible extent in the areas of morphology, syntax, and logical consistency. Initiate extensive research to develop new methods for performing semantic analysis and synthesis, establishing factual consistency or inconsistency, and resolving factual ambiguities. (U)

3. Adaptive Learning and Correlation Techniques

Initiate extensive research to develop appropriate deductive and inductive techniques for the automated processing of linguistically transformed input sentences in order to correlate their content. The investigation should include at least the following methods and means of implementation:

(a) Classification and association of data.

(b) Relevance of relationships.

(c) Reinforcement of learning.

(d) Data organization and structuring.

(e) Methods of retrieval.

(f) Logic and strategy of man-machine questioning.

(g) Utilization of existing heuristic techniques of machine learning and adaptation.

(h) Development of new and modified heuristic techniques. (U)

4. Computer Programming

Initiate research to determine the applicability and utility of existing programming languages and the development of requirements for a problem-oriented language for fact correlation. Initiate research in advanced programming techniques such as the dynamic allocation of storage, multiprogramming, parallel programming, and self-organization among programming units or subprograms. Initiate work to determine the specific requirements for a Fact Correlation System executive control program. (U)

5. Equipment

Begin an analysis of equipment requirements for a Fact Correlation

System with emphasis on an appropriate matching of input equipment with processing and storage components based upon estimated user requirements. Planning for an initial operational capability system should be in terms of equipment that is currently available or in the late stages of development. This configuration would include microsecond access core storage and an automatic text reading machine. Whether auxiliary storage will consist of disc files or magnetic tapes should be determined from further study of user requirements and the analysis of equipment configurations. Research in component requirements per se should emphasize man-machine communication during processing. The formulation of advance research requirements depends upon the specific functions needed to perform the language transformation and adaptive learning tasks. The sole exception is the recommendation to develop a parallel random access memory that can retrieve information in a single memory cycle. (U)

DISTRIBUTION LIST

Recipient	Copies
*Commander, Rome Air Development Center Griffiss Air Force Base, New York Attention: RAAPT	10
*Commander, Rome Air Development Center Griffiss Air Force Base, New York Attention: RAALD	1
*Commander, Rome Air Development Center Griffiss Air Force Base, New York Attention: ROZMSTT	[illegible]
Headquarters, Electronic Systems Division Directorate of Technology Hanscom Air Force Base, Massachusetts Attention: Capt. LeRoy Ross, ESREE	3
Intelligence and Electronic Warfare Directorate Rome Air Development Center Griffiss Air Force Base, New York Attention: Mr. R. Weber, RAWIP	[illegible]

63-3-1613

UNCLASSIFIED

DEPARTMENT OF THE AIR FORCE
AIR FORCE RESEARCH LABORATORY (AFMC)

28 AUG 06

MEMORANDUM FOR: HAF/ICIOD
1000 Air Force Pentagon
Washington DC 20330-1000

FROM: AFRL/IF
26 Electronic Parkway
Rome NY 13441-4514

SUBJECT: Mandatory Declassification Review (MDR) Request, Case 05-MDR-053
(Your Memo, dtd 27 Sep 2005, Same Subject)

1. Per your request, I have reviewed and completed the Mandatory Declassification Review (MDR) of the documents titled Fact Correlation for Intelligence Analysis, Volume 1, Applied Research Plan and Fact Correlation for Intelligence Analysis, Volume 2, Analysis of Technical Problems, each dated 15 Dec 62, RADC-TDR-62-461, Vol. I and RADC-TDR-62-461, Vol. II respectively. Both documents were written by Federal Electronic Corporation of Paramus NJ.

2. Based on my review I concluded the following:

a. A classification change occurred on 31Dec1974 as both documents were downgraded from Secret to Confidential.

b. Major portions of the documents' text are obsolete and describe hardware, systems, and technologies which are over 43 years old.

c. After reviewing the subject documents it was determined that no parts in either document should remain CLASSIFIED. The disclosure of the contents of either document is not expected to cause damage to US national security and there were no reasons why the UNCLASSIFIED portions should not be released.

3. This review was performed in accordance with Executive Order 12958, as amended and both documents were DECLASSIFIED: August 25, 2006. Please contact me immediately if you need further information related to this matter.

Donald W. Hanson

DONALD W. HANSON, SES
Director, Information Directorate

2 Atch
1. AD 354604, Vol I (U)
2. AD 354615, Vol II (U)

UNCLASSIFIED

DEPARTMENT OF THE AIR FORCE
AIR FORCE RESEARCH LABORATORY (AFMC)

04 January 2007

MEMORANDUM FOR DTIC-OCQ
ATTN: Larry Downing
Ft. Belvoir, VA 22060-6218

FROM: AFRL/IFOIP

SUBJECT: Distribution Statement Change

1. The following documents (previously unclassified/limited) have been reviewed and have been approved for Public Release; Distribution Unlimited:

AD354604, "Fact Correlation for Intelligence Analysis, Volume 1, Applied Research Plan", RADC-TDR-62-461, Volume 1.

AD354615, "Fact Correlation for Intelligence Analysis, Volume 2, Analysis of Technical Problems", RADC-TDR-62-461, Volume 2.

2. Please contact the undersigned should you have any questions regarding this change notification. Thank you for your time and attention to this matter.

Kelly Caggwin

KELLY CAGWIN
STINFO Officer
Information Directorate